Evidence-driven appraisal of students' careers using process mining: a case study

Evidence-Driven Appraisal of Students' Careers Using Process Mining: A Case Study

Claudia Diamantini[1], Laura Genga[2], Alex Mircoli[3*] and Domenico Potena[1]

[1]Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Via B. Bianche 12, Ancona, 60131, Italy.
[2]Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Groene Loper 3 , Eindhoven, 5612 AE, The Netherlands.
[3]Department of Economic and Social Sciences, Università Politecnica delle Marche, P.le R. Martelli 8, Ancona, 60121, Italy.

*Corresponding author(s). E-mail(s): a.mircoli@univpm.it;
Contributing authors: c.diamantini@univpm.it; l.genga@tue.nl;
d.potena@unvipm.it;

Abstract

Today's universities are more and more focused on improving their educational programs and supporting their students throughout their academic journey. A key aspect of such an effort is understanding which factors contribute to poor students' performance. This research illustrates how educational process mining techniques can be used to effectively uncover success and failure factors in students' academic journeys through a case study at an Italian university. The research reveals patterns related to adherence to curriculum requirements, strategies for taking exams, and the influence of various factors, such as the number of exams passed in the first year on graduation timelines. These findings offer valuable insights for educational institutions that might be used to, e.g., implement support mechanisms to enhance students' overall success rates.

Keywords: Educational process mining, Curriculum Mining, Student performance analysis

1 Introduction

Today's universities strive to improve their educational programs and support their students through academic endeavours. This challenge is of particular importance for Italian universities, given that approximately 40% of students drop out without successfully finishing their studies and only 30% manage to graduate within a year following the standard duration of their degree programme [1].

In this scenario, it becomes essential for Universities to employ tools and metrics to assess study programmes. To this end, in recent years, the Italian Ministry of University and Research proposed the AVA (*ita:* Autovalutazione - Valutazione - Accreditamento, *eng:* Self-assessment - Evaluation - Accreditation) system[1] to enhance teaching processes in Italy. This system includes planning and evaluation sheets for teaching activities, along with standardized indicators for assessment. The primary objective of the AVA system is to evaluate the academic journeys of students in Italian universities and draw attention to critical situations. Within the framework of this study, our attention is directed towards the following AVA indicators employed for assessing the *outcome* of universities concerning the academic journeys of students.

1. **Early**: students who took their degree within the standard duration of the degree programme are considered early graduates (iC02 indicator). For a bachelor's degree, it is three academic years, i.e., three years and six months.
2. **One year late**: students who graduated within one year after the period above-mentioned (iC17 indicator).
3. **Late**: students that took more than one year beyond the normal duration.

While the mentioned indicators provide universities with an evidence-based means to evaluate their educational systems, they present an aggregated outlook of students' behaviours, offering limited assistance in comprehending potential reasons behind students' delays. To uncover possible obstacles for students, one has to understand how they progress during their studies and whether they can complete their courses within the intended timeframe.

This research showcases the application of *Educational Process Mining* (EPM) techniques to identify possible success and failure factors in students' academic journey in a real-world case study. EPM aims to uncover patterns and trends within educational data to understand how educational processes are carried out and identify improvement opportunities [2].

Our investigation falls within the "curriculum mining" branch in EPM. The objective of this branch is to scrutinize data pertaining to students' academic journeys, specifically the sequence of credit-bearing activity registrations. The aim is to derive valuable insights into the curricular choices made by students. Given the high degree of freedom students have in deciding when to take their exams in Italian Universities, we are especially interested in uncovering potential relations between different exam-taking paths and students' graduation

[1] https://www.anvur.it/attivita/ava/

time. To this end, we apply EPM techniques to i) compare students' careers to the *study program*, which represents the order in which exams should be taken according to the curriculum's coordinators, focusing on detecting differences between *early* and *late* students, ii) analyze the exam-taking process of both students' categories, involving all activities surrounding the preparation of an exam, and iii) determine whether and how students' progression in the first year impacts their graduation time. It is worth noting that goal ii) has been mostly neglected by previous work in curriculum mining, focused on information related to when a student passed a course. However, focusing on passed exams only provides a limited overview of the student's efforts in preparing the course (e.g., a student may attempt an exam multiple times before passing the course). We argue that this data can provide in-depth insights into students' preparation for a course and to which extent one or more courses act as bottlenecks hampering students' preparation for other courses.

This manuscript is an extension of a previous study ([3]). Here, we provide a more detailed description of the methodological steps we follow and their rationale. Furthermore, we have significantly extended the predictive modelling analysis employed for goal iii). First, we explicitly modelled the impact of taking an exam later than when suggested by the study program on students' graduation time. Furthermore, we evaluated several classification algorithms, while only logistic regression was considered in the previous version. This analysis provided valuable insights into exam-taking patterns commonly associated with a higher or lower likelihood of students graduating on time.

The rest of this manuscript is organized as follows. Section 2 describes the case study and the research design. Section 3 discusses the obtained results. Section 4 provides an overview of relevant related work, while Section 5 draws some conclusions and delineates future work.

2 Study design

2.1 Case study: bachelor program of an Italian University

Our study focuses on a 3-year Bachelor's Degree program from an Italian university. As discussed in Section 1, the goal of this study is to uncover possible success and failure factors in students' academic journey in the given case study. We are particularly interested in investigating the impact of choices made by students regarding the order in which exams are taken, as previous studies have already suggested that these may impact their academic performance [4, 5]. This subject can, in principle, be investigated from different perspectives and with different data mining or machine learning techniques (see Section 4 for an overview). However, we argue that process mining techniques provide an effective means to deal with the high variability characterizing students' careers. In the considered case study, students have at least 7 opportunities per exam spread across the academic year, and the relative order exam opportunities of different courses are provided usually varies during the same or between different academic years. As a result, students' careers are likely to

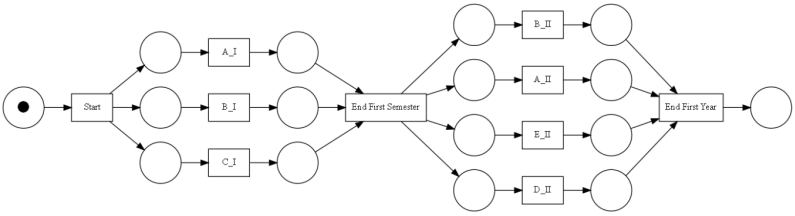
be very heterogeneous among each other, making it challenging for traditional data mining techniques for sequence analysis to infer relevant patterns. Process mining techniques, on the other hand, effectively detect different control-flow patterns (e.g., parallelism or choices) from such heterogeneous sequences, thus revealing regularities in students' behaviours that would likely remain hidden when considering the raw sequences of exams taken. This allows us to make a more robust distinction between exams (mostly) taken in a precise order, likely indicating some actual dependency, and exams (mostly) parallel, among which no clear dependency can be highlighted. In this paper, we aim to answer the following questions:

- **Question 1:** Are students following the study program more likely to graduate on time? Previous studies suggested that adherence to the study program could play a role in academic performance [4, 6]. We are interested in determining to what extent this adherence can be considered a success factor for our population.
- **Question 2:** What are typical behaviors for the preparation of the "critical" courses in the curriculum? This question aims first at determining which exams are potential blocks in students' careers, meaning they struggle to take the exam. Such exams are particularly important to detect since they will likely harm students' academic performance. Furthermore, we are interested in deriving the students' preparation process for these exams and determine whether there are noticeable differences between early and late students.
- **Question 3:** How does the students' progression in their first year impact the final grade and the graduation time? This question is meant to investigate whether assessing the risk of delay at the early stages of students' careers is possible. These insights would be valuable in determining which students are at risk and developing strategies to support them.

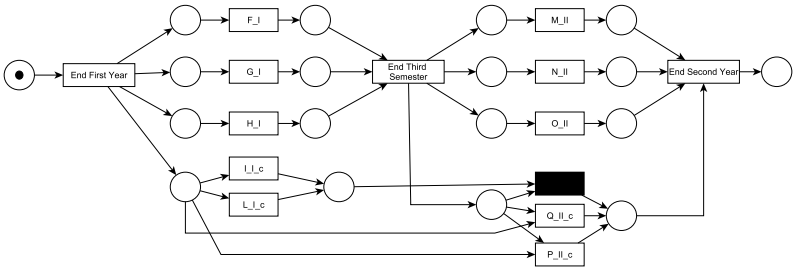
These research questions were particularly designed to give insights to a school manager, i.e., curriculum coordinator, faculty dean or department head. In fact, we believe that the answers to these questions can improve the understanding of how students approach their academic studies.

We analyzed the administration documents to derive the study program for each bachelor year of the case study under consideration. Figures 1a and 1b[2] show the manifesto for the first and second year processes, respectively, in *Petri net* notation. Places are graphically represented by circles and transitions by boxes. Places with tokens, represented as black dots inside a place, are possible *states* of the process, while labeled transitions correspond to process activities. Places represent possible *states* of the process, while labeled transitions correspond to process activities. For privacy reasons, we used the X_Y anonymization convention, where X is a progressive letter identifying the name of the course and Y identifies the semester. Courses with a logical sequence

[2]High-resolution versions of the figures shown in this paper are available at https://github.com/a-mircoli/edupm



(a) Study program: first year



(b) Study program: second year

Fig. 1: Study programs for the first and the second year

have the same letter X, e.g., Physics 1 and Physics 2. Second-year courses identified with an additional letter 'c' represent elective courses that can be chosen by the students. In detail, in addition to the six mandatory second-year courses, the student must choose one from the four elective courses, of which two are offered in the third semester and two in the fourth one. Note that a Petri net can also involve transitions without labels, visually represented as black boxes (see, e.g., Figure 1b). These transitions are referred to as *invisible* transitions, i.e. transitions that are not observed by the information systems and are mainly used for routing purposes. Edges model the order in which activities have to be executed. In details, an activity can be executed if each input place contains a token. If enabled, an activity consumes a token from each input place and produces a token for each output place. Finally, two paths are in parallel in the Petri net representing a process model if both must terminate before process execution can proceed.

The dataset contains 355 students who graduated in the academic years 2015/2016 to 2018/2019. We considered only these years to ensure that all the students in our sample were not affected during their first year of study by the effects of the COVID-19 pandemic, including changes in teaching delivery and examination methods. We acknowledge, however, that the COVID-19 has impacted the last years for some students in our sample. Since a good percentage of the students enrolled in 2017/2018 and 2018/2019 managed to graduate either on time or one year late, we make the assumption that the COVID-19 was not the main factor determining the performance of the late students.

Academic Year	Graduated Students	Early	One year late	Late
2015/2016	98	40.82 %	35.71 %	23.47 %
2016/2017	89	44.94 %	30.34 %	24.72 %
2017/2018	80	58.75 %	32.50 %	8.75 %
2018/2019	88	72.73 %	27.27 %	0.0 %

Table 1: Indicators students enrolled from 2015/2016 to 2018/2019.

Future studies should, however, be carried out to quantitatively validate this assumption.

The dataset was preprocessed with domain experts to remove outliers and wrong samples. More precisely, we removed from the dataset students' careers that were not possible according to the study program due to either some logging mistakes or to some special situations (e.g., students moving from one Bachelor's Degree program to another). The numbers displayed in Table 1 correspond to the preprocessed dataset used in the analysis. Table 1 shows that around the half (53.54%) of the graduated students managed to complete their studies on time, which suggests that the current setting of the program may involve important bottlenecks, thus making this a suitable case study for our analysis. It can be seen from Table 1 that there are very few or no students who graduated late in the cohort of those enrolled in 2017/2018 and 2018/2019, respectively. This results from having extracted the data at the beginning of 2023.

For each student, we have the year of enrollment, the overall duration of the graduation process in days and the grade for each exam the student passed. Moreover, for each exam that the student booked for, we also have information about the status and its timestamp. In this paper, we consider the following statuses: *Booked, Passed, Failed, Absent, Withdrawn*. Note that "Absent" means that the student has booked the exam but did not show up; while "Withdrawn" means that the student showed up but decided to withdraw. To gain a better understanding of when activities are carried out w.r.t. the academic year, we have inserted in the log four artificial activities as a time reference to indicate the end of each semester and year: "End first semester", "End first year", "End third semester" and "End second year".

2.2 Methodology

The following subsections delve into the approach adopted for each question.

2.2.1 RQ1: Students' career process analysis

The first question determines whether adhering to the proposed study program increases students' likelihood of graduating on time.

The analysis focuses on processes in the first and second years, as the third year offers more flexibility in class choices. Consequently, an individual's performance is influenced by their personal choices rather than the structure of the study program. For RQ1, the event log contains a trace of each student's

activities corresponding to passed exams. We are interested in quantifying the overall students' career performance and in pointing out which parts of the actual process deviate from the expected one. In particular, we want to understand the difference between the categories of early and late students. To address the first point, we employ *cost-based conformance checking* techniques [7]. This technique is state-of-the-art in quantifying the overall degree of compliance between a process model representing the normative behaviour and a set of log traces tracking the actual process executions. In particular, we leverage the *fitness* metric, which is a widely used metric ranging between 0 and 1 that quantifies to which extent the behaviors that are observed in the log comply with the behaviors allowed by the model. This notion of compliance is at the event level, meaning that the compliance level is computed for each trace separately. The more events that comply with the normative control flow of the model, the higher the compliance level of the trace is. Similarly, the more events that occur in a not-allowed position or are skipped, the lower the compliance level of the trace. The log fitness is an average of the trace fitness values. Then, to derive more in-depth insights, we delve into the actual students' careers. In particular, we employ *process discovery* techniques to derive a process model describing the *actual* behaviours of early and late students to gain insights on which process behaviours from the study program are followed or violated by the two categories of students. We used the *Infrequent Inductive Miner* (IM) algorithm [8], a process discovery technique commonly used in literature that can cope with infrequent behaviour and large event logs while ensuring soundness. The IM is required to set a threshold determining how much process behavior should be filtered out. We tested a range of thresholds from 0 to 0.9, selecting the best trade-off between fitness and precision (i.e., the ability of the model to generate only sequences belonging to the event log). We used the conformance checking and IM implementations provided in the Pm4Py library[3], an open-source library for process mining developed for the Python language. Note that, since the manifesto of the study program changed in 2017, we analyzed the conformance rate for students enrolled before and after that date separately.

2.2.2 RQ2: Exam taking process

Here, we are interested in determining which exams of the curriculum are "critical" exams, meaning that students usually struggle in taking them. These exams are interesting for our analysis since if their preparation takes a long time it can affect the preparation of other exams and ultimately lead to delays. In this sense, they can potentially become "bottlenecks" for the students. To identify critical exams, ideally, we would like to quantify the actual efforts put into the *preparation* of these exams. While direct observation of the preparation process is not possible, we can leverage information about how many times a student enrolled for an exam and the corresponding outcome before succeeding as a proxy for the preparation needed for that exam. To this end,

[3] https://pm4py.fit.fraunhofer.de/

we created an event log for each analyzed exam, where the student identifier is used as case id, the status correlated to each exam is used as activity, with the corresponding timestamp. Activities represent the path each student follows from first booking the exam to passing it: "Booked", "Passed", "Failed", "Absent", "Withdrawn". Activities, except passed, can occur several times consecutively. We first derive some simple statistics to identify exams showing a strong incidence of failure status. Then, we used the IM again to extract the exam preparation process. We chose to use process discovery (PD) for several reasons. First, PD allows us to obtain a more comprehensive and compact view of all possible process variants, while also highlighting the flow structures. Additionally, using IM, we can highlight only the most significant phenomena, excluding rare behaviors without excluding rare traces. Clearly, a disadvantage is obtaining process models that might be more complex to analyze, unlike methods such as variant analysis where individual variants are easier to analyze.

2.2.3 RQ3: Prediction

The goal of this analysis consists in assessing the impact of the students' progression during the first semester/year on their final graduation performance. A common practice to determine the impact of a set of features on a target variable is to leverage a predictive model to learn the input-output relation of interest and assess each feature's importance. We modelled the students' progression using features related to the exams taken, their grades, and whether or not each exam was taken on time. Unlike other works in the literature, we did not represent the process instances using encoding techniques, such as index-based encoding, as they add too many features and, considering the size of the dataset at hand, would lead to the curse of dimensionality problem. The target variable is a binary variable modelling students' graduation performance, i.e., whether they graduate on time. We tested multiple classification algorithms commonly used in literature to determine the classifier that best performs in our case study. In particular, we evaluated seven classification algorithms, namely Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), k-Nearest Neighbor (k-NN), Gradient Boosting (GB) and Long Short-Term Memory (LSTM). Among all the existing deep learning architectures, we considered LSTM due to its ability to deal with data sequences. Except for the LSTM network, we implemented all the algorithms using the *scikit-learn* library,a popular open-source machine learning library for the Python programming language that provides a wide range of tools for data mining and data analysis[4]. The LSTM network was implemented through the *Keras*[5] framework, which offers libraries for building classifiers based on both machine and deep learning architectures. To evaluate the classifier performance, we leverage the *accuracy* metric, which is a widely used metric in classification task, defined as follows. Let x_{ij} be the number of

[4] https://scikit-learn.org/stable/
[5] https://keras.io/

Log	#Traces	# Variants	Min E.p.T.	Max E.p.T.	Avg. E.p.T.
Early, 1st year	80	72	4	9	7
Late, 1st year	45	29	2	6	4
Early, 2nd year	80	80	6	12	9
Late, 2nd year	45	44	3	9	6

Table 2: Number of traces, number of variants and minimum, maximum and average number of events per trace (E.p.T.) of the first and second year for early and late students

Year	Fitness Early	Fitness Late
1st	0.74	0.59
2sd	0.70	0.44

Table 3: Fitness values for students enrolled in 2015/2016 and 2016/2017

samples belonging to j-th class which have been classified as i-th class. Let C be the number of classes and N be the total number of data. The accuracy achieved by a classifier is computed as: $\frac{1}{N}\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C} x_{ii}$. Two other commonly used metrics are *precision* and *recall*, which respectively measure the ratio of the number of correctly classified samples of a class to the total number of samples predicted as or really belonging to the same class. The validation of the models was carried out through the *10-fold cross-validation*, which is a widely used technique in machine learning. The dataset is divided into 10 equal parts (folds). The model is trained and tested 10 times, each time using a different fold as the test set and the remaining 9 folds as the training set. Accuracy achieved on each model are then averaged to provide a reliable estimate of the model's performance.

3 Results

3.1 Students' career processes

In the following, we discuss the results obtained for students enrolled in 2015/2016 and 2016/2017. We obtained similar trends for the early students of the second group of students; however, we only have 7 late students for the students enrolled in academic years 2017/2018 and 2018/2019, from which no representative model could be extracted.

Table 2 describes the event logs of early and late students' careers for the first and the second year. The numbers confirm the expected high variability in students' behaviors, as can be seen by the high number of different process variants for both students' categories. This is especially true in the second year, due to optional exams. Table 3 shows the fitness values achieved for early and late students enrolled in the academic years 2015/2016 and 2016/2017, for the first and the second year. In both cases, early students show a much higher level of compliance than late students, which suggests the higher the conformance

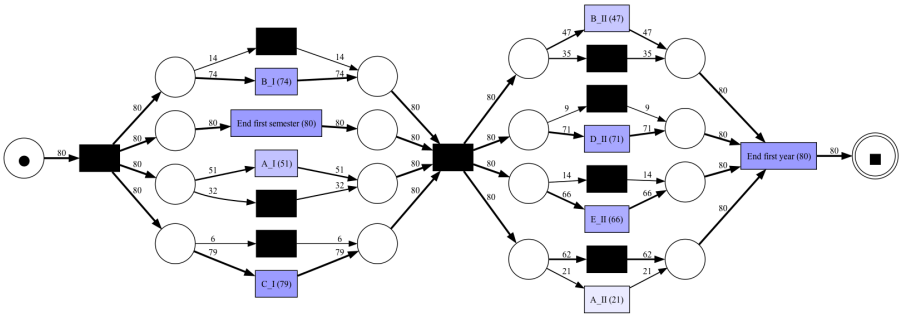


Fig. 2: Model extracted for the first year of the Early students

to the study program, the higher the likelihood that the student will graduate on time. Figure 2 shows the model extracted for the early students. Edge labels correspond to the number of students following the corresponding path. Note that some numbers appear not to sum up correctly due to filtering. For example, considering exam B_I, the model shows that 74 students took it, while 14 did not; however, there are 80 students in total, so this result appears contradictory. This is because some students took the exam after one of the exams of the second semester. Since this is an infrequent behaviour, the iIM does not report it in the model. As a result, it sums up on the edge of the silent activity modelling skipping the exam B_I both the students not taking the exam in the first year and those taking it in a different stage.

A first interesting observation is that the process model extracted for early students resembles the study program presented in Figure 1a. Recalling the meaning of parallelism in our context introduced in Section 2, this model suggests that most of the early students take exams in the first semester before exams of the second semester. There are, however, exceptions; indeed, all parallel exams can be skipped at the beginning of the process, except "End first semester". The fact that "End first semester" is in parallel with other exams indicates that some students manage to take them before the end of the first semester, while others don't. Note that this is likely due to some exam opportunities offered after the end of the first semester but before the start of the exam sessions for the second semester. Overall, the model suggests that most early students pass most of the first year exams within the first year, whose transition is sequential to the second examination block in parallel. Delving into the activity frequencies, almost all students pass B_I and C_I within the first semester, while slightly fewer students pass A_I. In the second semester, only 26% of students pass A_II, which is the advanced level of A_I.

The model extracted for late students is shown Figure 3a. We can notice that even though the first part of the model presents some similarities with the study program, the overall control-flow differs quite significantly from it, especially after the first semester. Looking at the path in parallel with the "end of first semester" transition, it follows that there are some students who do not take exams in the first semester and this is a quite a frequent behavior

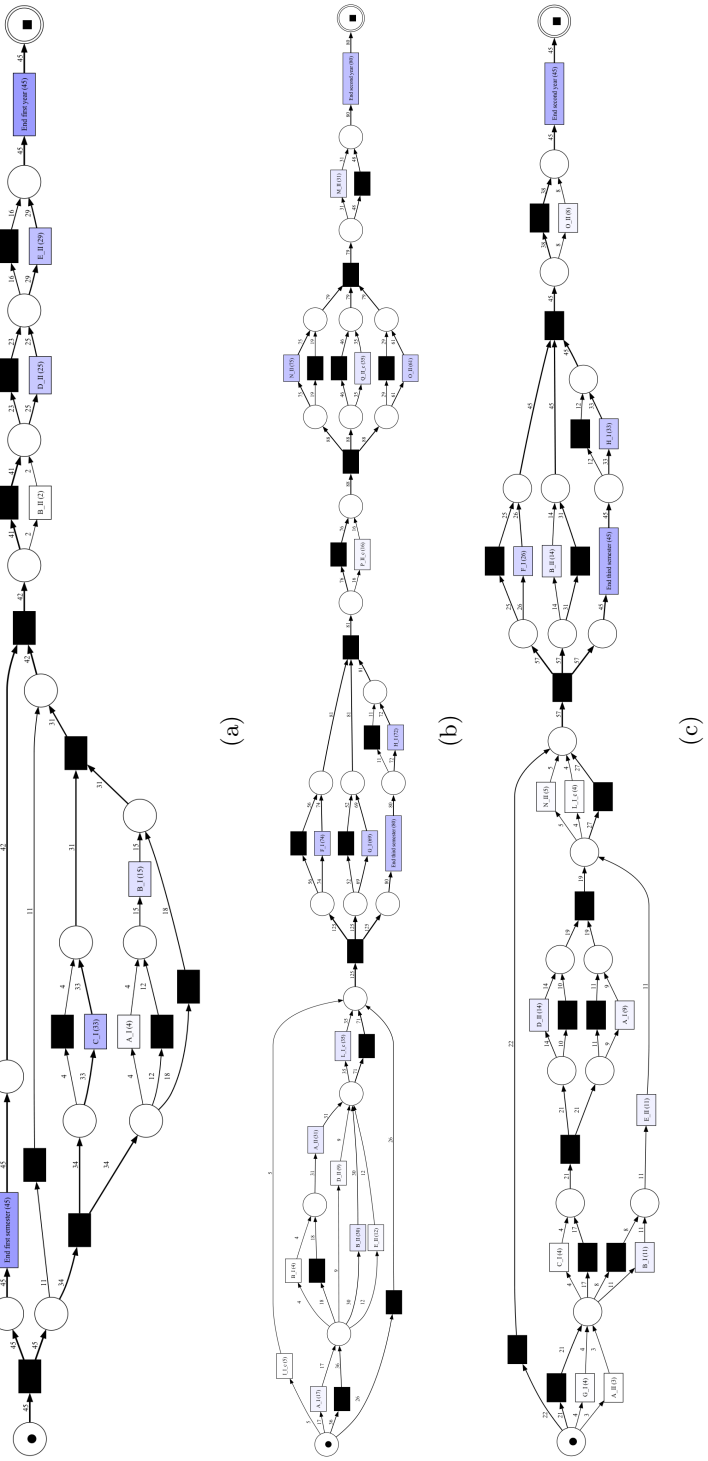


Fig. 3: Models extracted for the first year of late students (a) and for the second year of early (b) and late (c)

(indeed, it occurred for a quarter of the students). While C_I is still modelled as parallel to A_I and B_I, an ordering relation is now detected among the latter, indicating that students who took A_I usually prefer to take it before B_I. The model presents escape paths to take only one of the first semester exams, similar to what we observed for early students. However, for late students, the pass rate of all exams in the first semester drops dramatically, especially for A_I, from 63.75% (for early students) to only 8.89%. Furthermore, we can notice that A_II is not reported in the model, which means that none (or very few) students took this exam in the first year. After the first semester, no parallelism is detected, indicating a clear tendency for these students to focus first B_II, then on D_II and then on E_II. The reasons for such a preference might vary and should be discussed with domain experts. For instance, the students might decide to take first the exams perceived as more complex to be able to leverage additional opportunities or, on the contrary, the ones perceived as the easiest. At the same time, the model also shows us that many students did not take either of these exams. In particular, students who pass B_II are far less than those who pass D_II and E_II.

Regarding the models for the second year, shown in Figure 3b and Figure 3c, the differences are less evident. In both processes, there are still exams that students did not manage to pass in the first year. However, early students show a tendency to take the first-year exams before the end of the third semester, with the result that while the first part of the discovered model is quite chaotic and involves many different paths, the last semester is quite close to the reference model. The model for the late students shows however a different trend, where B_II can be taken after the third semester and where the majority of the students do not manage to take the exams of the fourth semester by the end of the second year. Notably, few of the late students pass A_I by the second year, confirming that it is a course to focus on. M_II would also seem difficult to pass. In fact, it is found only in the process model related to early students, while less than 40% of students passed it. G_I, N_II and O_II drop sharply moving from early to late students model, i.e., from 55.20%/85.23%/69.32% to 8.00%/10.00%/17.78%, respectively. Finally, about half of the students passed F_I e H_I.

To summarise, exam A_I is potentially a critical exam since fewer students take it, both for early and late students. Consequently, also A_II represents an issue. In the same way, exam M_II can be considered critical. If we consider only the late students, also B_I, B_II, G_I, N_II and O_II exams appear to be potentially critical exams. We delve into the analysis of critical exams in the next section.

3.2 Exam taking process

The previous analysis highlighted some exams which are frequently taken by the students later than expected. However, these results do not tell us whether the students decided to postpone the exam preparation at a later stage in their career or, instead, they started to prepare on time but encountered challenges

Exam	Absent/Success	Failed/Success	Withdrawn/Success
A_I	68.89 %	166.17 %	4.2 %
B_I	75.99 %	133.66 %	11.39 %
C_I	41.38 %	0 %	12.81 %
A_II	74.5 %	80.75 %	2.25 %
B_II	93.63 %	21.57 %	17.89 %
D_II	9.56 %	0 %	14.46 %
E_II	74.15 %	0 %	0 %
F_I	20.78 %	34.96 %	9.05 %
G_I	6.22 %	0 %	37.06 %
H_I	4.39 %	231.22 %	0.24 %
M_II	47.75 %	12.5 %	5.75 %
N_II	27.32 %	147.07 %	10.49 %
O_II	16.23 %	13.25 %	11.59 %

Table 4: Percentage of absent, failed and withdrawn student for each exam

in succeeding. The goal of this section is to gather more detailed insights on which exams are actual bottlenecks, in the sense that their preparation takes much longer than expected, possibly hampering the preparation of other exams.

Table 4 shows, for each exam, the percentage of *absent*, *failed* and *withdrawn* students, out of the total number of students who passed the exam (i.e., 191 students). Note that cell values may exceed 100% because a student may be absent, withdrawn or fail the exam even several times before passing.

For our analysis, in accordance with domain experts, we consider an exam to be "critical" if it has a percentage of students that have failed them over 100 %. Critical exams are highlighted in grey in Table 4. For the first year, these are A_I and B_I. Other exams showing concerning trends are A_II and B_II, since many students registered for the exam but chose not to attend. In the second year, we can observe a critical situation for exams H_I because the failure rate is greater than 200 %, and N_II with a failure rate of 147.07 %.

To delve into students' preparation process for these exams, we mined the process models corresponding to the taking process of the detected critical exams, then compared the obtained models for early and late students. For the sake of space, in the following, we only show the results obtained for the processes of one of the detected critical exams, i.e., A_I.

Figure 4 represents the process of students who graduated on time taking the exam A_I. This process shows three parallel paths. In the first path, we can see a loop of students that are absent, suggesting that they studied for the exam but they did not think they were prepared enough to pass it. However, this loop can also be skipped: in fact, there are students who manage to pass the exam the first time they take it. The second path adds the time perspective to the analysis: half of the students (53.33%) pass the exam by the first year, 28.33% by the end of the third semester, and the rest by the end of the second year. In the last path there is a loop for the students who took the exam but failed. The process ends with the "Passed" activity for all the students.

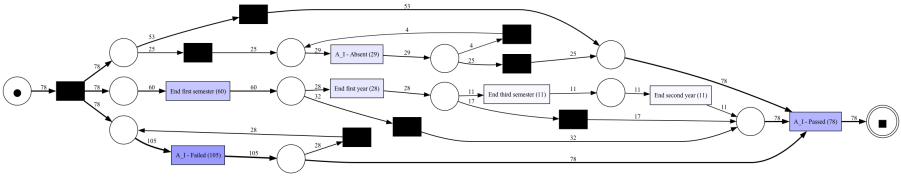


Fig. 4: A_I Early

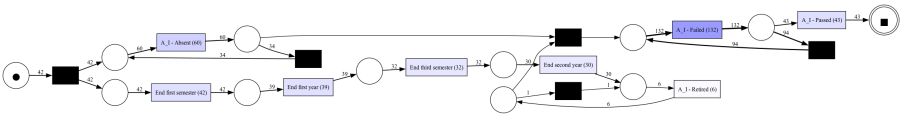


Fig. 5: A_I Late

Figure 5 represent the same process but refers to students who did not graduate on time. Many more unprepared students do not show up or fail the exam, compared to the model in Figure 4. Moreover, most students pass the exam after the end of second year.

The analysis of the two models shows that, as expected, early students take exams from the first sessions. In fact, there are few absentees and many take the exam within the first semester (even if they fail it). We derive that such students start studying as early as the course progresses. As for late students, more than 71% take the exam after the end of the second year. Considering that if a student books for an exam, has started or plans to start studying it, then it follows that late students have longer preparation time.

3.3 Predicting students' outcome

This subsection addresses the third RQ, aimed at understanding to which extent students' progression in their first semester and first year impacts their graduation performance. Since the changes in the study manifesto did not impact the first year, we considered the whole event log for this analysis. The analysis presented in Section 3.2 shows that many students, especially *late* ones, do not manage to take first-year exams within the time window as defined in the study program. Hence, we were interested in uncovering potential relations between the exams passed by students during the early stages of their university careers, together with their grades, and the final students' outcomes. The choice to also consider the grades is motivated by the fact that a high grade can certainly mean that the student is good, but it can also indicate a long time to study at their best. Furthermore, we observed that many students take exams multiple times, possibly to obtain a higher grade, spending more time preparing for an exam, which in turn may lead to some delay in their graduation.

We performed an extensive phase of parameter tuning through grid search by varying parameters' values in commonly used ranges and then choosing

Algorithm	D1	D2	D3
DT	0.749	0.772	0.792
RF	0.769	0.803	0.806
SVM	0.766	0.792	0.758
NB	0.741	0.741	0.724
K-NN	0.744	0.769	0.792
GB	0.772	0.789	0.811
LSTM	0.769	0.791	==

Table 5: Best average accuracy achieved by each classification algorithm. Best results for each dataset are highlighted in gray.

the values that gave the best results, in terms of the highest average accuracy obtained using the 10-fold cross validation. In particular, the chosen ranges for each algorithm are as follows:

- DT: max_depth ∈ {2, 3, 4, 5}; criterion ∈ {"entropy", "gini", "log_loss"}; min_samples_split ∈ {2, 3, 4, 5}; min_samples_leaf ∈ {1, 2, 3};
- RF: n_estimators ∈ {10, 20, 50, 100, 200}; criterion ∈ {"entropy", "gini", "log_loss"}; min_samples_split ∈ {2, 3, 4, 5}; min_samples_leaf ∈ {1, 2, 3};
- SVM: C ∈ {0.1, 1, 10, 100, 1000, 10000, 100000}; kernel ∈ {"linear", "poly", "rbf", "sigmoid"}; class_weight ∈ {"None", "balanced"}; max_iter ∈ {-1, 10, 50, 100, 500, 1000};
- k-NN: n_neighbors ∈ {3, 4, 5, 6, 7}; weights ∈ {"uniform", "distance"}; metric ∈ {"cosine", "minkowski"};
- GB: loss ∈ {"exponential", "log_loss"}; n_estimators ∈ {20, 50, 100, 200}; learning_rate ∈ {0.001, 0.01, 0.1, 1}; max_depth ∈ {2, 4, 6, 8, 10};
- LSTM: lstm_units ∈ {10, 20, 50, 100}; epochs ∈ {10, 20, 50, 100}.

For what concerns the Naïve Bayes algorithm, no optimization of the parameters was carried out as there were no tunable parameters.

First, we considered two datasets in which grades taken in the first semester (D1) and in the first year (D2), respectively, were considered as features. D1 has 3 features, one for each exam of the first semester, and D2 has 7 features, i.e., all the exams of the first year. Each feature has been normalized linearly so that the minimum value of each feature took the value 0 and the maximum the value 1. For each sample in the two datasets, a label has been added to indicate whether the student belongs to the class of students who graduated on time (*Early*) or the class of those who graduated one or more years late (*Late*). For each algorithm, the best results in terms of average accuracy are shown in the first two columns of Table 5.

For the dataset D1, the best result is obtained using GB (accuracy achieved is 0.772), while all other classifiers have slightly lower performance, between 0.741 and 0.769. Including information regarding the second semester (D2) allows us to obtain results that are 2% higher on average than those obtained on the D1 dataset for each algorithm, with the best accuracy achieved using RF (0.803). Again, the performance of the various algorithms is quite similar, with a variance of 3.69E-04. It should be noted that, to take advantage of LSTM peculiarities, the two datasets have been enriched with information about the

	Actual Late	Actual Early	Precision
Predicted Late	137	38	**0.783**
Predicted Early	29	151	**0.839**
Recall	**0.825**	**0.799**	

Table 6: Confusion matrix of the best result for D3

order in which the exams were passed. However, no improvement in accuracy was found. A possible explanation of the phenomenon can be identified in the low cardinality of the considered dataset. It is likely that by extending the dataset by considering additional academic years and/or degree programs, better results could be obtained.

It is noteworthy that, except for the LSTM network, the results for the dataset containing the first-year grades (D2), were obtained by including features that do not take into account the semester in which the exam was taken. Indeed, an exam in the first semester could have been taken in both the first and second semesters, or never taken during the first year. However, the order in which exams are taken could affect the prediction of a student's career, as some exams of the first semester are preparatory to those of the second one. For this reason, we built an additional dataset (D3) obtained from D2 by multiplying by -1 the grade of the first semester exams that were passed in the second one. The results are reported in the third column of Table 5. For D3, Gradient Boosting turns out to be the best classifier, with a slight improvement with respect to the best result of dataset D2 (i.e., accuracy=0.811). All results are slightly higher than those obtained with the other datasets, except for SVM and NB, whose performance is slightly worse. Note that we did not test LSTM on D3 because the sequence with which the first-year exams were taken was already considered for this algorithm in the D2 dataset.

The confusion matrix of the best model for D3 is shown in Table 6.

From the confusion matrix, it can be seen that good precision and recall results are achieved for both classes. These metrics are especially important for the Late class. Indeed, low precision results in wasted economic resources and time to activate initiatives to support students who will not actually graduate late. On the other hand, a low recall represents a limit for the model in identifying students who may have difficulty graduating on time.

To interpret results, considering that the differences in accuracy are minimal among the models shown in Table 5, we report in Figure 6 the decision tree obtained from the whole dataset. To read the model correctly, it should be noted that, in the Italian university system, an exam is considered passed if the grade ranges from 18 to 30. It turns out that passing A_II is a key indicator in determining whether a student will graduate on time. If A_II is passed within the first year, the student will belong to the Early class with an accuracy of about 98%. Students will be classified as Early (with 96% accuracy) if they pass at least one between A_II and B_II (about 21% of students). In combination with the results presented in Sections 3.1 and 3.2, this output offers interesting insights. Indeed, the previous analysis revealed that A_I and B_I

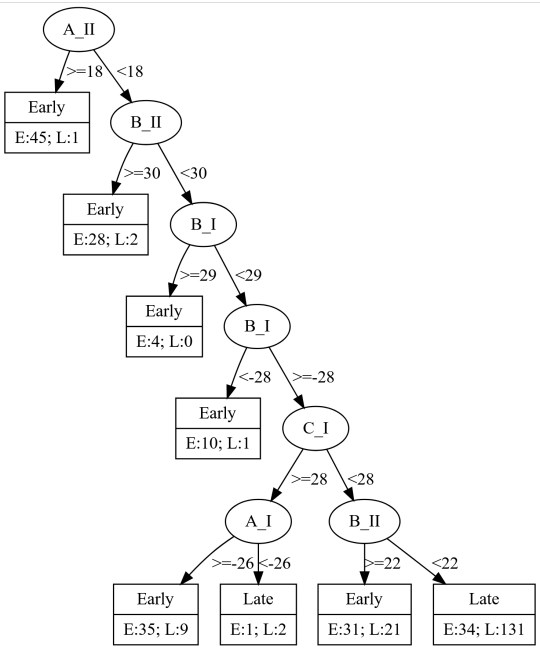


Fig. 6: Decision Tree for D3

represent a bottleneck for most students, with a high failure rate. One would expect that succeeding or failing these exams impacts when students are ready to take their direct successors, i.e., A_II and B_II. Hence, indirectly, we would expect to see a similar impact on the overall students' performance. However, this analysis shows that the impact of taking A_I and B_I on time is much lower than the impact of passing A_II and B_II on time. Passing B_I alone does not affect prediction significantly; in fact, the decision is made only if the student has a grade greater than 29 (remember that a grade with a negative value indicates that a first-semester exam was passed in the second semester). In other cases (B_I with a grade less than 29 or failed), the prediction depends on C_I, A_I, and B_II. A student will be classified as Late (with an accuracy of about 80%), only if they have failed A_II and B_II within the year, C_I has a grade under 28 (or failed) and B_II has a grade under 22 (or failed).

4 Related work

Educational Data Mining (EDM) is an emerging discipline that aims to understand and improve students' learning process to enhance educational outcomes, support personalized learning, and provide valuable insights for educators, administrators, and policymakers in education ([9], [10]). Numerous methods, accompanied by empirical studies assessing their efficacy, have been suggested to tackle various tasks. Some of the key tasks include: creating social networks

that depict students' interactions in e-learning activities ([11], [12]); grouping students with similar learning patterns or behaviors, to identify common characteristics or needs ([13], [14]); text mining and sentiment analysis to process students' feedback to understand their opinions and insights ([15]); provide the students with tailored recommendations, e.g., on the courses to enroll ([16]); mining sequential patterns describing students' learning behaviors or progression ([17]); prediction of students' performance using historical data ([18], [19]). Our work is mainly related to EDM approaches that analyze students' academic performance and their failure. A popular trend in this respect consists in modeling students according to predefined features and applying machine learning to predict student's performance ([20], [21], [22], [23], [24], [25]). Many of these studies provide a perspective complementary to the one provided by our analysis, taking into account factors external to the graduation process itself. Even studies centred on the graduation process usually perform a data-oriented analysis, in which students' behaviors are encoded in terms of features without considering the study program's underlying structure. In this respect, our work is similar to the one in [26], which proposes to model and analyze students' careers. They present the concept of an "ideal career," which represents the career trajectory of a graduated student who completed each exam immediately after concluding the corresponding course. As certain exams may be taken in the same semester, an expert is employed to establish the optimal order for taking exams, resulting in a sequence where each exam is distinguished by its position. The career of each student is then represented as a sequence of integers, with each integer indicating the position the exam should have occupied according to the ideal progression. Different metrics measure the distance between each student's career and the ideal one (e.g., the Bubblesort distance). They also utilize sequential pattern mining methods to deduce the most prevalent subsequences of exams. Each element within the sequence corresponds to either exams taken in the same semester or those taken with a certain delay, measured in terms of semesters. Compared to our approach, the work in [26] does not exploit the potentialities of process-based analysis in modelling students' behaviors. In contrast, we exploit process formalisms and *process mining* techniques to explicitly model possible parallelism among exams, allowing us to evaluate the difference between single careers and the ideal path more accurately.

The application of process mining techniques to educational data, referred to as *Educational Process Mining* (EPM) ([2]), is a subject that has been recently gaining increasing interest. EPM has been applied to deal with different educational problems, such as on-line learning environments ([27], [28], [29]), computer-supported collaborative learning tools ([30], [31]), professional training ([32], [33]), self-regulated learning assessment [34]. However, only a few works investigated the applications of EPM to curriculum mining. [35] propose a set of patterns modeling typical constraints of academic curricula and use these patterns to analyze the graduation process. However, unlike

the present work, they do not infer the process model representing students' careers and do not focus on delay analysis.

An alternative approach is proposed in [36], where the authors apply conformance-checking techniques to analyze students' curricula.

Our approach is similar to [37], in analyzing the exam-taking process and program study compliance. However, [37] do not consider differences between early and late students.

The authors in [38] propose to distinguish between successful and unsuccessful students but they do not evaluate their compliance with the manifesto. Moreover, the paper only gives a general overview of the approach, without providing a true experimental evaluation. Also in [39] the authors take into account the division between successful and unsuccessful students instead, but their approach involves using simulated data to create a recommendation system that suggests students which exam to take next.

In [40], authors apply process mining techniques to analyze event log data generated within educational information systems, to understand students' behavior during online learning. The work differs from ours in two main ways: (i) it is based on the concept of digital twin to represent students' activities and (ii) its focus is on the single course while ours is on the entire career. To the best of our knowledge, only a few works consider the entire student's career. [41] focus on changes in students' learning behaviors over time. Each semester, they extract a student profile describing the number of exams given at the right moment, anticipated, postponed, and repeated, together with performance indicators such as the grade average. These profiles are then used to cluster students, and cluster evolution analysis techniques are employed to detect changes in cluster characteristics over time. The output of this study is complementary to ours, which instead aims to extract a process model describing the orders with which students took the curricula exams. [5] proposes to model students' trajectories as sequences of *backpacks*, i.e., sequences of failed exams that the students have to retake. Directly Follows Graphs are used as modelling formalism, where each node represents the set of failed exams and edges are used to denote transitions from one backpack to the other. Our study employs a different perspective as we focus on passed exams. The study from [42] presents some similarities to ours since they investigate how to apply the PM^2 methodology to understand students' path and analyze their conformance to the suggested path adopting a process perspective. However, they do not make a distinction between successful and late students, and they do not focus on analyzing bottlenecks in the study program. [6] applies process discovery techniques to curriculum event logs to characterize behaviors of students that performed best/worst in terms of years required to complete the graduation process and final grade. A similar approach is implemented in a recent study from Diamantini et al. [4]. In this work, they shift the focus to classes of students defined according to the indicators defined by the Italian Minister of Education. Moreover, they investigated students' compliance with the manifesto and the students' delays in taking their exams. Our study adopts the same

classes of students, and share some of their goals, e.g., in terms of assessing the compliance of students' careers to the manifesto. However, in contrast to [4], we also analyzed the exam-taking process to highlight patterns associated with early and with late students. Furthermore, we investigated the predictive value of students' career process features for predicting students' performance. We argue that identifying relations between students' careers and outcomes provides valuable insights into generating actionable recommendations for newly enrolled students on structuring their graduation process.

5 Conclusion and future works

This study has delved into the landscape of students' career processes. Process mining techniques allowed us to map out students' journeys. Our exploration of early and late students' behaviors concerning the curriculum revealed distinctive patterns and potentially blocking trend, notably revolving around specific exams such as A_I and A_II. These exams impact students' graduation times, making them key areas for attention and intervention. We also delved into the exam preparation process for the identified critical exams, uncovering insightful patterns in how the two analyzed categories of students usually prepare for these exams. Despite the high variability in students' careers, PM techniques were able to identify typical patterns in the progression of early and late students, allowing us to filter out outliers and irrelevant behaviors to obtain a model of the core process of each student's category. Furthermore, PM techniques explicitly discover and model concurrency, which is a crucial characteristic in the university domain. Indeed, students usually prepare for multiple exams in the same period and differences in terms of exam sequence are usually due to administration choices related to the exam dates, rather than a real dependency among two subjects. Finally, we applied machine learning algorithms to evaluate the impact of the number and the grade of exams taken in the first year on the overall students' graduation performance. We are considering to use the results of this work as a proactive tool to early detect cases of delayed graduation and try running corrections by providing students with additional tools such as tutors, additional exercises, and so on.

Nevertheless, some limitations should be considered when evaluating the study results. First, the models describing the students' careers (RQ1) only consider the date an exam has been passed. While this provides us with some insights on which exams are prepared close to each other, little can be said about potential concurrency within the preparation of different exams. In future work, we intend to investigate the feasibility of leveraging the information related to the exam preparation process (used in RQ2) to derive at least an approximation of which exams are usually prepared together. Furthermore, the interpretation of these process models is sometimes not straightforward, as shown by the unexpected placing of the synthetic activities *end of first semester* and *end of first year*. Finally, it should be noted that the employed discovery techniques return lossy models, which do not cover all the students'

behaviors. In future work, we plan to investigate the application of local model discovery techniques to identify relevant portions of process behaviors that might remain hidden when start-to-end models are employed.

In conclusion, this research encourages future analysis to improve the quality of education by predicting the students' academic performance and supporting those in the risk group. In future work, we plan to extend the students' samples by taking into account different courses to determine to which extent we can determine similar or different patterns. We also intend to investigate the use of variant analysis techniques to determine groups of students characterized by more homogeneous behaviors and assess their academic performance. Furthermore, we intend to investigate how to convert the insights derived from this kind of analysis into actionable recommendations that can support the students in determining the best path to follow at different moments of their careers, keeping into account their current progression.

Declarations

5.1 Ethical Approval

Not applicable

5.2 Availability of supporting data

Supporting data are not available

5.3 Competing interests

The authors have no competing interests as defined by Springer, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

5.4 Funding

This research has received funding from the project Vitality – Project Code ECS00000041, CUP I33C22001330007 - funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - 'Creation and strengthening of innovation ecosystems,' construction of 'territorial leaders in R&D' – Innovation Ecosystems - Project 'Innovation, digitalization and sustainability for the diffused economy in Central Italy – VITALITY' Call for tender No. 3277 of 30/12/2021, and Concession Decree No. 0001057.23-06-2022 of Italian Ministry of University funded by the European Union – NextGenerationEU.

5.5 Authors' contributions

C.D.: Conceptualization, Methodology, Supervision, Writing - Review & Editing L.G.: Methodology, Software, Writing - Original Draft, Writing - Review

& Editing A.M.: Software, Writing - Original Draft, Writing - Review & Editing D.P.: Conceptualization, Methodology, Supervision, Writing - Review & Editing

5.6 Acknowledgments

Not applicable

References

[1] C. Aina, F. Pastore, Delayed graduation and overeducation in italy: A test of the human capital model versus the screening hypothesis. Social Indicators Research **152**(2), 533–553 (2020)

[2] A. Bogarín, R. Cerezo, C. Romero, A survey on educational process mining. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(1), e1230 (2018)

[3] D. Potena, L. Genga, A. Basta, C. Mercati, C. Diamantini, *Evidence-Based Student Career and Performance Analysis with Process Mining: A Case Study*, in *International Conference on Process Mining* (Springer, 2023), pp. 349–360

[4] C. Diamantini, L. Genga, A. Mircoli, D. Potena, N. Zannone, Understanding the stumbling blocks of italian higher education system: A process mining approach. Expert Systems with Applications **242**, 122747 (2024)

[5] J.P. Salazar-Fernandez, J. Munoz-Gama, J. Maldonado-Mahauad, D. Bustamante, M. Sepúlveda, Backpack process model (BPPM): A process mining approach for curricular analytics. Applied Sciences **11**(9), 4265 (2021)

[6] M. Cameranesi, C. Diamantini, L. Genga, D. Potena, *Students' careers analysis: a process mining approach*, in *Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics* (ACM, 2017)

[7] A. Adriansyah, B.F. van Dongen, W.M. van der Aalst, *Conformance checking using cost-based fitness analysis*, in *2011 IEEE 15th International Enterprise Distributed Object Computing Conference* (IEEE, 2011), pp. 55–64

[8] S.J. Leemans, D. Fahland, W.M. Van Der Aalst, *Discovering block-structured process models from event logs containing infrequent behaviour*, in *Business Process Management Workshops: BPM 2013 International Workshops* (Springer, 2014), pp. 66–78

[9] A. Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works. Expert Systems with Applications **41**(4), 1432–1462 (2014)

[10] C. Romero, S. Ventura, Educational data mining and learning analytics: An updated survey. Wiley interdisciplinary reviews: Data mining and knowledge discovery **10**(3), e1355 (2020)

[11] K.L. Cela, M.Á. Sicilia, S. Sánchez, Social network analysis in e-learning environments: A preliminary systematic review. Educational Psychology Review **27**, 219–246 (2015)

[12] S. Yassine, S. Kadry, M.A. Sicilia, Detecting communities using social network analysis in online learning environments: Systematic literature review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(1), e1431 (2022)

[13] S. Križanić, Educational data mining using cluster analysis and decision tree technique: A case study. International Journal of Engineering Business Management **12**, 1847979020908675 (2020)

[14] E. Trandafili, A. Allkoçi, E. Kajo, A. Xhuvani, *Discovery and evaluation of student's profiles with machine learning*, in *Proceedings of Balkan Conference in Informatics* (ACM, 2012), pp. 174–179

[15] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, L. Galligan, Sentiment analysis and opinion mining on educational data: A survey. Natural Language Processing Journal **2**, 100003 (2023)

[16] S.B. Aher, L. Lobo, Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. Knowledge-Based Systems **51**, 1–14 (2013)

[17] J. Wong, M. Khalil, M. Baars, B.B. de Koning, F. Paas, Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course. Computers & Education **140**, 103595 (2019)

[18] W. Xiao, P. Ji, J. Hu, A survey on educational data mining methods used for predicting students' performance. Engineering Reports **4**(5), e12482 (2022)

[19] C. dos Santos Garcia, A. Meincheim, E.R.F. Junior, M.R. Dallagassa, D.M.V. Sato, D.R. Carvalho, E.A.P. Santos, E.E. Scalabrin, Process mining techniques and applications–a systematic mapping study. Expert Systems with Applications **133**, 260–295 (2019)

[20] G.W. Dekker, M. Pechenizkiy, J.M. Vleeshouwers, Predicting students drop out: A case study. International Working Group on Educational Data Mining (2009)

[21] S. Gowda, R. Baker, Z. Pardos, N. Heffernan, *The sum is greater than the parts: Ensembling student knowledge models in ASSISTments*, in *Proceedings of KDD Workshop on Knowledge Discovery in Educational Data* (2011)

[22] H. Guruler, A. Istanbullu, M. Karahasan, A new student performance analysing system using knowledge discovery in higher educational databases. Computers & Education **55**(1), 247–254 (2010)

[23] S. Herzog, Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. Research in Higher Education **46**(8), 883–928 (2005)

[24] G. Lassibille, L. Navarro Gómez, Why do higher education students drop out? Evidence from Spain. Education Economics **16**(1), 89–105 (2008)

[25] C. Romero, S. Ventura, P.G. Espejo, C. Hervás, *Data mining algorithms to classify students*, in *Proceedings of International Conference on Educational Data Mining* (2008), pp. 8–17

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

[27] A. Bogarín, C. Romero, R. Cerezo, M. Sánchez-Santillán, *Clustering for improving educational process mining*, in *Proceedings of International Conference on Learning Analytics And Knowledge* (ACM, 2014), pp. 11–15

[28] P. Mukala, J.C. Buijs, M. Leemans, W.M. van der Aalst, *Learning Analytics on Coursera Event Data: A Process Mining Approach*, in *Proceedings of International Symposium on Data-Driven Process Discovery and Analysis* (2015), pp. 18–32

[29] J.C. Vidal, B. Vázquez-Barreiros, M. Lama, M. Mucientes, Recompiling learning processes from event logs. Knowledge-Based Systems **100**, 160–174 (2016)

[30] R. Bergenthum, J. Desel, A. Harrer, S. Mauser, in *Transactions on Petri Nets and Other Models of Concurrency V* (Springer, 2012), pp. 22–50

[31] P. Reimann, J. Frerejean, K. Thompson, *Using process mining to identify models of group decision making in chat data*, in *Proceedings of*

International Conference on Computer Supported Collaborative Learning (International Society of the Learning Sciences, 2009), pp. 98–107

[32] R. Bergenthum, J. Desel, A. Harrer, S. Mauser, *Learnflow mining*, in *6th e-Learning Fachtagung Informatik (DeLFI)* (2008), pp. 269–280

[33] A.H. Cairns, B. Gueni, J. Assu, C. Joubert, N. Khelifa, *Analyzing and improving educational process models using process mining techniques*, in *Proceedings of International Conference on Advances in Information Mining Management* (2015), pp. 17–22

[34] R. Cerezo, A. Bogarín, M. Esteban, C. Romero, Process mining for self-regulated learning assessment in e-learning. Journal of Computing in Higher Education **32**(1), 74–88 (2020)

[35] N. Trcka, M. Pechenizkiy, *From local patterns to global models: Towards domain driven educational process mining*, in *Proceedings of International Conference on Intelligent Systems Design and Applications* (IEEE, 2009), pp. 1114–1119

[36] Y. Bendatu, B.N. Yahya, Sequence matching analysis for curriculum development. Jurnal Teknik Industri **17** (2015)

[37] R. Hobeck, L. Pufahl, I. Weber, *Process Mining on Curriculum-Based Study Data: A Case Study at a German University*, in *International Conference on Process Mining* (Springer, 2022), pp. 577–589

[38] R. Wang, O.R. Zaiane, Discovering process in curriculum data to provide recommendation. Educational Data Mining (EDM) 2015, (2015)

[39] R. Wang, O.R. Zaïane, *Sequence-based approaches to course recommender systems*, in *Database and Expert Systems Applications* (Springer, 2018), pp. 35–50

[40] A. Azeta, F. Agono, F. Adesola, V. Nwaocha, S. Tjiraso, A process mining framework for analysing students' behaviours using digital twin. Available at SSRN 4331450 (2022)

[41] S.A. Priyambada, M. Er, B.N. Yahya, T. Usagawa, Profile-based cluster evolution analysis: Identification of migration patterns for understanding student learning behavior. IEEE Access **9**, 101718–101728 (2021)

[42] R. Hobeck, L. Pufahl, I. Weber, *Process Mining on Curriculum-Based Study Data: A Case Study at a German University*, in *Process Mining Workshops* (Springer Nature Switzerland, Cham, 2023), pp. 577–589