Understanding the stumbling blocks of Italian higher education system: A process mining approach

Claudia Diamantini [a], Laura Genga [b], Alex Mircoli [a,*], Domenico Potena [a], Nicola Zannone [c]

[a] *Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy*
[b] *Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, Eindhoven, The Netherlands*
[c] *Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands*

ARTICLE INFO

ABSTRACT

Nowadays universities strive to continuously enhance their educational programs to improve both the quality and quantity of their graduates. This is a sensitive problem, especially for Italian universities where only 30% of the students enrolled at the university succeed in graduating within a year after the normal duration of the study plan. Over the last few years, the Italian Ministry of University and Education has introduced several indicators to assess students' careers and help universities identify possible criticality in their study programs. However, these indicators only provide a high-level overview of the graduation process without providing insights into students' failure. To address this issue, in this work, we propose to model a study program as a process and exploit process analysis techniques to assess students' performance. These techniques allow delving into students' careers, thus enabling the investigation of their failures and delays. The findings obtained by applying our approach to the Bachelor program of an Italian university allowed us to determine common bottlenecks that seem to have an impact on students' graduation time. Moreover, we were able to determine and compare the career paths of successful and late students. The insights gathered by our analysis can be used to support university personnel in delving into factors causing some exams to be a bottleneck, as well as to determine potential improvements in the overall curricula.

1. Introduction

Stumbling blocks encountered by students during their university careers are a major challenge of high-level education, as they are the main cause of delayed graduation. This phenomenon is particularly critical for Italian universities, as reported in Aina and Pastore (2020), as around 40% of the students fail to complete their studies and only 30% graduate within a year after the normal duration of their study plan. The main causes are linked to: (i) the traditional choice of universities to define unconstrained paths for students, which is too free and, given that there are no constraints on when to take exams, means that exams can be taken without having taken first the suggested exams, and (ii) the uneven preparations of the students, due to the lack of entrance tests in many universities. In such a context, it is crucial for universities to have tools and metrics to identify possible causes of students' failures. In recent years, the Italian Ministry of University and Education proposed the AVA (*ita:* Autovalutazione - Valutazione - Accreditamento, *eng:* Self-assessment - Evaluation - Accreditation) system[1] for the improvement of Italian teaching, which

provides planning and evaluation sheets for teaching activities, as well as standard indicators for evaluation. The AVA system has the aim of assessing students' careers in Italian universities and highlighting critical situations. Some examples are the percentage of students who acquire a given number of academic credits within the first year of enrollment, the percentage of students who complete their studies within the predefined time window, the percentage of students who quit their studies, and so on (see Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, 2021 for a complete list of indicators). Although these indicators provide universities with useful insights into students' academic performances, they provide a high-level, aggregated view of students' behavior, which provides little support in identifying the causes of students' failures and delays and, hence, in determining possible strategies to solve them. To identify courses representing potential blocks for students, one has to delve more into how students proceed during their studies, i.e. whether they are actually able to take their courses in the expected time window.

* Correspondence to: Via Brecce Bianche, 22 - Ancona (AN), Italy.
 E-mail addresses: c.diamantini@univpm.it (C. Diamantini), l.genga@tue.nl (L. Genga), a.mircoli@univpm.it (A. Mircoli), d.potena@univpm.it (D. Potena), n.zannone@tue.nl (N. Zannone).

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

In the last decade, there has been an increasing interest towards *Educational Data Mining* techniques aimed at analyzing data generated during students' educational processes to gain insights into students' learning behaviors, with the aim of, for instance, determining factors affecting whether the students complete their studies (e.g., students' gender, background, average grades), or providing tailored recommendations to students or teachers to improve the effectiveness of a given learning process (Peña-Ayala, 2014; Sanjeev & Zytkow, 1995).

In this work, we aim to provide universities with a tool to investigate students' performance and gain useful insights into which courses represent a block for students. In particular, we implement principles of the EDM discipline in a real case study, i.e. students from a Bachelor program at an Italian University, to determine bottlenecks affecting students' performance in terms of time needed to complete a graduate program. In particular, our analysis belongs to the so-called "curriculum mining" branch of EDM, whose goal consists in analyzing data related to students' *careers*, i.e. the sequence of registrations of credits-bearing activities by the university in the transcript of records of individual students, to determine valuable insights on the curricula chosen by students. We adopt a *process perspective* to model a study program and students' careers. We model the "ideal" career (hereafter referred to as the "manifesto"), i.e. the order in which students are supposed to take the exams according to the study program, as a process; the *careers* of individual students are modeled as process instances. Based on this representation, we investigate the application of *process mining* principles to detect possible bottlenecks in students' paths and identify root causes of delays. Process mining aims to explore logs recording process executions to distill valuable knowledge about the corresponding process (van der Aalst, 2011) and has been recently applied to facilitate the understanding of educational processes (Bogarín et al., 2018). We apply process mining techniques to identify typical students' careers and assess their adherence to the manifesto, which is supposed to help students acquire the knowledge necessary to complete their studies step-by-step successfully. The results show that careers compliant with the manifesto are more likely to be associated with successful students (i.e., students that graduate within one year after the end of the degree program). Therefore, the most frequent stumbling blocks for students are caused by the non-compliance of their careers to the manifesto, in the sense that taking an exam before having taken the exams suggested before it (e.g., Calculus II before Calculus I) involves greater difficulty in understanding the subject and, consequently, in passing the exam. As a consequence, comparing the actual students' careers with the manifesto allows us to identify discrepancies that can point out possible problems encountered by students. Such findings can help universities implement a career monitoring system to evaluate the implemented improvement tasks and support students during their careers. The study also outlined the need for additional class hours for the courses that resulted in being more difficult for students and helped the university define new teaching programs.

Outline. The remainder of the paper is organized as follows. Section 2 presents motivations for our work and introduces the case study. Section 3 describes the procedure and techniques we applied to our case studies. Section 4 discusses the results of our experiments. Finally, Section 5 discusses related work, and Section 6 draws some conclusions and directions for future work.

2. Motivations

The Italian Ministry of Education, Universities and Research requires Italian universities to assess the quality of their education on a yearly basis by means of a set of indicators (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, 2021; Ministero dell'Università e della Ricerca, 2019) that are part of the AVA system. In this work, we focus on the indicators commonly used to evaluate the *output* of universities in terms of students' careers. The Ministry distinguishes three main classes of students' careers:

- *Successes*: career of students that successfully graduated within one year after the end of the degree program (Indicator iC17)
- *EarlyFailures*: career of students that dropped off the degree program within the first year (Indicator iC14)
- *LateFailures*: career of students that dropped off the program within one year after the end of the degree program excluding early failures (Indicator iC24 minus iC14)

Note that these indicators do not completely describe the entire student population. There might be students that take more than one additional year to graduate or students that drop out after the fourth year. However, they are considered of less interest by university stakeholders.

While the indicators above provide universities with a means to evaluate their education system, they do not provide insights into the root causes of students' failures and delays. To this end, universities often perform an internal assessment of their study programs to identify possible blocks for students and, thus, carry out countermeasures to improve their study programs. This evaluation is typically exam-driven. In particular, students' performance for a given academic year is assessed by computing the percentage of exams successfully passed by students in that year.

Next, we illustrate the indicator-based and exam-based analysis using a Bachelor program at an Italian University as a case study. Note that, in our analysis, we only consider *mandatory* courses. This is motivated by the fact that these are typically the most critical exams in a study program. Indeed, students are required to pass them in order to graduate, and failing them has likely a negative impact on their careers. Moreover, focusing on mandatory courses allows us to analyze students' careers with respect to exactly the same paths, ensuring a fair comparison between exams and between student groups.

2.1. Case study

For our study, we consider a 3-year Bachelor Degree program from an Italian university. Fig. 1 shows the manifesto of the study program. Due to privacy reasons, data are anonymized using the following convention: XY_Z, where X is a progressive capital letter identifying the course name, Y represents the year, and Z the semester. For courses that are the continuation of other courses, such as Calculus I and Calculus II, we added the '_b' suffix to the second course. It is worth noting that each academic year is divided into two semesters and that mandatory courses are only related to the first two years.

Table 1 shows the indicators computed for the program for students enrolled from academic years 2011–2012 to 2014–2015, on the basis of data extracted in December 2022. Hereafter, for the purpose of this study, we consider only students who successfully graduated or dropped out of the degree program. Indicators score similarly for the different academic years, with some peaks as regards success students in 2013 and early failure students in 2012 and 2014.

Table 2 shows students' performance for academic years from 2011–2012 to 2014–2015. The worst performance values (i.e., success ratio below 50%) are highlighted in bold. Students seem to encounter challenging exams already in the first year. The average success ratio for first-year courses in 2011 is 43.07%; also, three courses out of seven seem to represent a potential block, i.e. "A1_1", "A1_2_b", "C1_2_b". An overall improvement can be observed for first-year courses in the academic year 2012–2013; the average success ratio is 48.81% with two potential blocks, although with better results compared to the ones observed in 2011–2012. Performance looks much worse for second-year courses: with the exception of "F2_1" and "H2_1", all exams achieved a success ratio below 50%. In 2013–2014, results look quite close to those obtained in 2012–2013, with the difference that in this year also "A1_1" has a success ratio below 50% and, on the opposite, "M2_2" registered a significant improvement. Similar results were also obtained in 2014–2015.

The analysis of the Bachelor program shows an overall improvement in students' performance over the years. This might be due to several
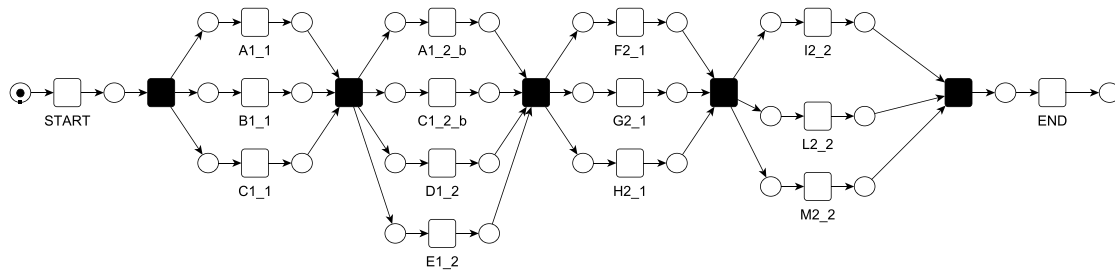
Fig. 1. Manifesto for mandatory courses of the considered case study.

Table 1
Indicators related to the students enrolled in 2011, 2012, 2013, and 2014. The number of students is reported in parentheses.

Academic year	#Enrolled students	Successes (iC17)	EarlyFailures (iC14)	LateFailures (iC24 - iC14)
2011–2012	130	27.69% (36)	1.54% (2)	14.61% (19)
2012–2013	121	24.79% (30)	8.26% (10)	8.26% (10)
2013–2014	123	34.95% (43)	1.63% (2)	10.57% (13)
2014–2015	137	16.79% (23)	8.13% (10)	9.76% (12)

Table 2
Success ratio for exams taken within academic years 2011–2012 to 2014–2015.

Course	2011–2012		2012–2013		2013–2014		2014–2015		2015–2016	
	#Enr. Stud.	Success Ratio (%)	#Enr. Stud.	Success Ratio (%)	#Enr. Stud.	Success Ratio (%)	#Enr. Stud.	Success Ratio (%)	#Enr. Stud.	Success Ratio (%)
A1_1	130	**36.15**	121	50.41	123	**47.15**	137	**42.34**		
B1_1	130	62.31	121	66.94	123	82.11	137	75.91		
C1_1	130	**49.23**	121	66.94	123	64.23	137	51.82		
A1_2_b	130	**16.92**	121	**24.79**	123	**30.08**	137	**37.23**		
C1_2_b	130	**12.31**	121	**33.06**	123	**46.34**	137	62.77		
D1_2	130	57.69	121	58.68	123	82.11	137	99.27		
E1_2	130	59.23	121	97.52	123	61.79	137	90.51		
F2_1			128	50.78	111	53.15	121	83.47	131	71.76
G2_1			128	**29.69**	111	**33.33**	121	**41.32**	131	**39.69**
H2_1			128	57.81	111	83.78	121	83.47	131	79.39
I2_2			128	**17.19**	111	**32.43**	121	**43.80**	131	63.36
L2_2			128	**23.44**	111	**37.84**	121	**35.54**	131	**43.51**
M2_2			128	**45.31**	111	72.97	121	59.50	131	64.12

factors, such as a change in the responsible teachers, significant changes in the course topics or in the exam. Nevertheless, delving into the data it turns out that these differences are partially due to the metric used by universities to assess students' performances. Since the students' enrollment year is not taken into account when computing the number of students that passed a course in a given year, we might actually include in this sum students from previous years, i.e. students that took the course late.

Such an issue points out how an exam-oriented analysis, although providing useful hints to identify potential blocking exams, can over-estimate students' success ratio, thus affecting the reliability of the assessment. In particular, one could miss some exams that actually represent a block for students because the success ratio looks in line with other exams because of the (hidden) counting of students from previous years, which is clearly not desirable. To overcome this drawback, we propose to adopt a *process-oriented* perspective for the analysis of students' careers, as described in the following section.

3. Methodology

In this work, we aim to provide universities with a tool to investigate students' performance and gain useful insights into which courses represent a block for students. In particular, we aim to answer the following research questions:

Q1: What are the common bottlenecks in students' careers?

Q2: What is the actual students' career?

Q3: To what extent does the adherence of students to the manifesto impact their performance?

Q4: What are the main differences between the careers of success and late-failure students?

To be able to answer such questions, we have to shift the focus of the analysis from exams, as currently done by Italian universities, to students' careers. A key to enabling the analysis of students' careers is to take into account the academic year in which students enrolled in a study program. This allows us to understand the order in which students typically give their exams and grasp useful insights on critical issues and blocks, which would otherwise remain hidden when analyzing the indicators proposed to assess the quality of education or using an exam-based analysis (see Section 2).

In the remainder of the section, we present the datasets used for our analysis and the methodology adopted to answer the questions above.

3.1. Datasets

For our analysis, we extracted data concerning students' careers and, in particular, student enrollment and exam registration from
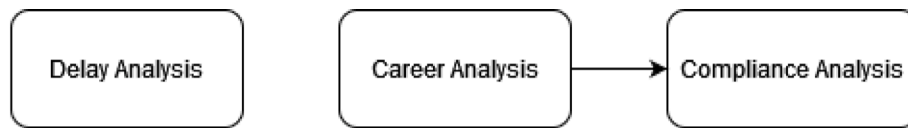
Fig. 2. The proposed methodology for the analysis of students' careers.

ESSE3. ESSE3 is a software suite widely used by Italian universities to provide online educational services to students and staff. For each student, we extracted the ID (after proper anonymization), the enrollment year, and the graduation or withdrawal year, when available (at the time of data extraction many students were still enrolled in the study program). Then, for each student, we extracted the set of exams she took, each described by the exam name, the scheduled year, the scheduled semester, and the date on which the student passed the exam.

A preliminary analysis of the gathered datasets revealed that some students graduated without meeting the study program's requirements. For instance, 16 students present some anomalies in their careers. Specifically, some exams were recorded before the students could have followed the corresponding course, according to the enrollment year. These situations likely represent students who moved from a different study program at the same university or from another university. We removed those students from our datasets as they are not representative for our analysis.

3.2. Approach

To answer the research questions posed above, we take advantage of recent developments in the field of *process mining*. Process mining is a broad discipline that encompasses a plethora of techniques to turn data stored in an information system into valuable insights on the underlying *processes* (van der Aalst, 2011). Roughly speaking, a process consists of a set of *activities* that have to be performed by some *actors* to reach a certain goal. A *process instance* (also called *case*) is a specific execution of the process, recorded by an information system in an *event log* in the form of *traces*, i.e. sequences of *events* recording the execution of activities in their order of occurrence.

In our context, we model the manifesto of a study program as a process where activities correspond to the courses that students have to take in order to successfully complete their studies (cf. Fig. 1). Students' careers, representing the set of exams taken by a student, are modeled as process instances. To this end, we need to move from the relational model in which data are stored in ESSE3 to an event log. Accordingly, we extracted information related to exams taken by each student and we used the student ID as the case ID. The registration date of the exams was used to determine the order of events in each trace. We also added two artificial events to each trace to represent the beginning (*START*) and the end (*END*) of the trace. Moreover, we added to each trace four artificial events to represent: *(i)* the end of the first semester, *(ii)* the end of the first year, *(iii)* the end of the third semester, and *(iv)* the end of the second year. These events are needed to add a temporal dimension to the models extracted in the career analysis (see below) and, thus, enable more accurate analysis.

In the remainder of the section, we present the approach used to answer research questions from **Q1** to **Q3**. In particular, we identify three different phases, as shown in Fig. 2. Each related to a different research question: *Delay Analysis*, *Career Analysis*, *Compliance Analysis*. The first step is independent from the others, while the second and the third steps are subsequent. We address these questions for *(i)* the entire set of students, *(ii)* success students, and *(iii)* late-failure students. This distinction allows us to address the research question **Q4**. Note that our analysis ignores the actual "effort" of individual courses (usually expressed in ECTS). We leave the analysis of students' careers based on ETCS for future work.

3.2.1. Delay analysis

A goal of our analysis is to identify which exams represent a block for students (**Q1**). To this end, we investigate whether students take their courses as expected. More precisely, for each course, we identify the students that took the exam within the same academic year in which they followed the course (hereafter referred to as *on-time* students). We also calculate the average time (expressed in days) required by students to pass the exam from the end of the semester in which the course was taught. In our analysis, we have considered January 1st as the end of the first semester and June 1st as the end of the second semester.

It is worth noting that this analysis is similar to the exam-based analysis described in Section 2. However, there are significant differences between the two analysis. While the exam-based analysis in Section 2 focuses on the exams taken by *all* students (regardless of the enrollment year), here the analysis focuses on the exam taken by students enrolled in a certain academic year. We argue that this shift of focus provides more accurate insights on possible blocks for students. In fact, this analysis makes it possible to spot the courses characterized by a *low* percentage of on-time students. Note that determining which percentage should be considered "low" depends on the context of the analysis. In this work, we set the threshold to 50%, i.e. all exams that were delayed by more than half of the target student group, are considered potential blocks.

3.2.2. Career analysis

The delay analysis provides an overview of students' performance with respect to single exams. The goal of the career analysis is to identify the actual students' careers (**Q2**), taking into account the order in which exams were taken by students. In particular, we apply *process discovery* to mine a process model representing the typical careers of students. Given an event log, process discovery techniques aim to construct a process model that provides an abstract representation of the underlying processes (see van der Aalst et al., 2003; van Dongen et al., 2009 for a review on process discovery techniques).

Given the high variability in students' behaviors, representing all careers would likely lead to chaotic and meaningless models. Therefore, we focus only on the most relevant trends of the process. To this end, we apply the *Heuristic Miner* algorithm (Weijters et al., 2006), which employs a set of heuristics aimed to filter out less relevant behaviors and reports in the model only the strongest causal dependencies. For our analysis, we used the heuristic miner implementation available in ProM,[2] which is an open-source framework for process mining. We refer to Weijters et al. (2006) for a detailed description of Heuristic Miner. Here, we only mention that we use Heuristic Miner in its default settings for what concerns noise thresholds. In our analysis, we exploit the option "All tasks connected" which guarantees that all process activities are shown in the model.

We visualize the models mined by Heuristic Miner using the *Causal net* (C-net) formalism (van der Aalst et al., 2011). This format is commonly used in process mining to model business processes. We have chosen this formalism since it is tailored to Heuristic Miner results and provides a more compact and simpler representation of the routing logic inferred from Heuristic Miner than the one obtained by using other modeling formalisms (e.g., Petri nets). A C-net is a directed graph,

---

[2] https://promtools.org/

Table 3

Delay analysis for students from 2011 to 2014. Column **Avg Time** indicates the number of days that, on average, students required to pass the exam from the end of the semester in which the course is given.

| | Enrolled students | | | | | | | |
| | 2011 | | 2012 | | 2013 | | 2014 | |
Exam	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time
A1_1	**36.1%**	413.7	**40.5%**	447.4	**36.6%**	334.5	**24.1%**	307.2
B1_1	62.3%	213.9	55.4%	239.4	58.5%	256.2	57.7%	198.9
C1_1	**49.2%**	221.3	55.4%	280.1	51.2%	290.8	**35.8%**	247.6
A1_2_b	**16.9%**	521.1	**13.2%**	638.3	**5.7%**	539.7	**4.4%**	475.4
C1_2_b	**12.3%**	608.4	**10.7%**	612.8	**21.9%**	389.8	**24.8%**	274.0
D1_2	57.7%	206.0	**43.0%**	224.9	**48.0%**	198.2	63.5%	115.2
E1_2	59.2%	185.3	72.7%	144.4	50.4%	213.6	65%	106.0
F2_1	50.8%	286.5	**42.3%**	347.7	60.3%	211.0	53.5%	149.0
G2_1	**29.7%**	490.0	**20.7%**	608.0	**21.5%**	478.5	**28.3%**	199.2
H2_1	57.8%	204.7	69.4%	205.3	70.2%	186.2	73.2%	178.6
I2_2	**17.2%**	501.5	**9.9%**	620.5	**15.7%**	391.9	**15.7%**	333.0
L2_2	**23.4%**	439.3	**21.6%**	474.5	**19.0%**	356.5	**29.1%**	126.3
M2_2	**45.3%**	260.8	51.3%	196.6	**39.7%**	228.7	**44.1%**	113.0

Table 4

Delay analysis for success students from 2011 to 2014. Column **Avg Time** indicates the number of days that, on average, students required to pass the exam from the end of the semester in which the course is given.

| | Success students | | | | | | | |
| | 2011 | | 2012 | | 2013 | | 2014 | |
Exam	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time
A1_1	83.3%	181.2	93.3%	137.7	65.1%	282	82.6%	183.6
B1_1	97.2%	81.5	96.7%	90.4	88.4%	131.2	82.6%	96.5
C1_1	88.9%	128.1	96.7%	96.1	88.4%	144.0	87%	130.0
A1_2_b	55.6%	273.0	**33.3%**	289.5	**13.9%**	433.5	**26.1%**	312.2
C1_2_b	**41.7%**	226.7	**40.0%**	313.7	**48.8%**	208	82.6%	93.3
D1_2	86.1%	93.4	76.7%	146.4	67.4%	99.1	91.3%	67.1
E1_2	91.7%	57.4	93.3%	69.7	67.4%	157.7	91.3%	47.2
F2_1	94.4%	81.1	63.3%	257.1	90.7%	122.6	91.3%	62.3
G2_1	77.8%	200.8	60%	289.9	53.5%	306.8	95.6%	98.0
H2_1	97.2%	87.3	96.7%	121.5	93%	145.5	100%	134.1
I2_2	55.6%	179.4	**36.7%**	289.9	**34.9%**	281.4	56.5%	136.6
L2_2	69.4%	156.8	70%	144.9	**48.8%**	180.3	82.6%	64.7
M2_2	88.9%	55.6	86.7%	81.1	65.1%	149.7	82.6%	69.8

where nodes represent activities and arcs represent causal dependencies between activities. Each activity has a set of possible input/output *bindings*, represented by dots placed on edges, which provide the routing logic of the control flow. More precisely, given an activity *a*, each output binding of *a* represents a set of activities that occurred concurrently after it; similarly, each input binding represents the (set of) activity(/-ies) that preceded activity *a*. Multiple bindings on the same edges represent 'OR' constructs.

3.2.3. Compliance analysis

The last step of our analysis aims to assess to what extent the adherence of students to the manifesto of the study program has an impact on their careers (**Q3**). To this end, we compare the ideal career, represented by the manifesto, against the actual students' careers.

To assess the degree of compliance of students' careers with the manifesto, we compute the *fitness* between the traces representing students' careers and the process model representing the manifesto. Fitness is a metric widely used in compliance analysis to quantify how much of the behavior observed in the event log is captured by the process model. The fitness value ranges between a minimum of 0 and a maximum of 1. To this end, we employ *alignment-based conformance checking* (van der Aalst et al., 2012). Conformance checking aims to determine possible mismatches between the observed and the intended process behavior. The notion of *alignment* provides a robust approach to conformance checking able to pinpoint the causes of nonconformity (see van der Aalst et al., 2012 for a formal definition of alignment). Intuitively, given a trace and a process model, an alignment maps the trace to a complete run of the model. In presence of deviations, some moves in the trace cannot be mimicked by the model or vice versa.

Fitness is computed based on the number of deviating moves in the alignment.

We perform the compliance analysis for all enrolled students as well as for the student groups identified by the ministerial indicators, namely students with a successful career and with a late-failure career. The aim is to assess whether adherence to the manifesto has an impact on students' careers. We apply hypothesis testing to assess the statistical significance of the differences between student groups.

4. Results

In this section, we present the result of our analysis of the Bachelor program presented in Section 2.1.

4.1. Delay analysis

Table 3 reports, for each exam, the percentage of students enrolled in a given academic year who managed to take the exam on time (**OnTime**) and the time (expressed in days) that, on average, students require to pass each exam (**AvgTime**). For column OnTime, values in bold correspond to exams for which the percentage of students on time was less than 50%. Table 4 and Table 5 report the results for success students and late-failure students, respectively.

We can observe that the performance of students enrolled in 2011 for first-year courses is the same that we observed in Table 2. This was expected since our dataset does not include students enrolled before 2011. However, it is worth noting that there are slight differences concerning the performance in the second year. In particular, the percentage of students on time is lower than the values we observed

Table 5
Delay analysis for late-failure students from 2011 to 2014. Column **Avg Time** indicates the number of days that, on average, students required to pass the exam from the end of the semester in which the course is given.

Exam	Late-failure students							
	2011		2012		2013		2014	
	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time	On Time	Avg Time
A1_1	**10.5%**	72.5	**10.0%**	196.0	–	–	–	–
B1_1	**42.1%**	182.8	**40.0%**	70.2	53.8%	204.0	**41.7%**	230.2
C1_1	21.0%	113.0	**30.0%**	84.0	**7.7%**	271.5	16.7%	150.5
A1_2_b	–	–	–	–	–	–	–	–
C1_2_b	–	–	0.0%	291.0	–	–	0.0%	162.0
D1_2	**47.4%**	129.8	50%	100.1	**15.4%**	109.3	**33.3%**	53.2
E1_2	**42.1%**	88.7	70%	57.3	**30.8%**	76.4	50%	144.0
F2_1	**5.3%**	175.0	**10.0%**	41.0	**7.7%**	19.0	**8.3%**	18.0
G2_1	**5.3%**	62.0	**0.0%**	–	–	–	**8.3%**	172.5
H2_1	21.0%	78.0	**30.0%**	162.0	**15.4%**	124	**16.7%**	144.0
I2_2	–	–	–	–	–	–	–	–
L2_2	–	–	–	–	–	–	–	–
M2_2	**10.5%**	83.0	**10.0%**	135.0	**7.7%**	47.0	**0.0%**	266.0

before for all courses. Delving into the data, we found that these differences are related to three students who were enrolled in 2012 but took second-year exams in the first year. This is likely due either to errors in recording the exam dates or to some exceptional circumstances (e.g., students moving from other study programs). We stress that, without taking into account the enrollment year, we cannot identify those cases.

Students from all academic years apparently encountered some challenges in both the first and second years of their studies. Some exams exhibit a low percentage of on-time students for all academic years, i.e. "A1_1", "A1_2_b", "C1_2_b", "G2_1", "I2_2", "L2_2" and "M2_2". These results are in line with those of Table 2; nonetheless, we found two additional exams, i.e. "D1_2" and "F2_1", that seem to represent a block for students, at least in some academic years, which on the other hand did not exhibit any issue using the exam-based analysis (Table 2).

In general, our analysis shows that the percentage of students who passed their courses on time is much lower than what is shown by the exam-based analysis. For example, the success ratio for "A1_2_b" in Table 2 is always around 30%, while here it is significantly lower, in some years below 10%. Similarly, "C1_2_b" registered a success ratio higher than 60% for the academic year 2013–2014 (Table 2); in contrast, our analysis shows that only 26.24% of the students actually managed to take it on time. The situation is much better for success students, where there are only a few exams that turned out to be a block (i.e., "C1_2_b" and "A1_2_b") and much worse for late-failure students, whereas, on the contrary, almost all exams were largely taken in delay. It is worth noting that some exams were not taken at all by late-failure students, i.e. "C1_2_b", "I2_2" and "L2_2".

Fig. 3 shows the average delays in days. For all academic years, we observe that approximately half of the mandatory courses were taken on time by a significant percentage of students. However, there are a few exceptions that seem to vary significantly from year to year. For example, the delay of students enrolled in 2012 for "A1_2_b" shows that, although most of the students delayed the exam of at least six months, there was still a relevant percentage of students able to pass it after a few months from the end of the course. On the other hand, among students enrolled in 2013, almost nobody managed to take the exam with a delay smaller than six months, suggesting that this exam turned out to be a critical block for students. An opposite trend can be observed for "L2_2", which has been taken by all students enrolled in 2013 with a delay lower than six months, while most students enrolled in 2012 show a larger delay, even more than one year. As expected, we obtain better trends when considering success students. Among all academic years, there are few exams for which we observe a delay larger than one year for this student group, e.g. "A1_2_b" and, for 2012, "F2_1" and "G2_1".

Finally, we observe that failure students usually took exams with a short delay. However, as discussed in the previous section, this result is mainly due to the low number of students in this group that passed mandatory courses.

4.2. Students' career

We now analyze the typical careers of enrolled students. Table 6 shows the characteristic of the event log used to infer students' careers. Note that we removed two students from the dataset, one of which belonging to the late-failure group, since these students did not take any mandatory course during their studies.

Figs. 4, 5, and 6 show the models obtained for all three student groups by applying Heuristic Mining, while Fig. 7 shows the graphical notation used to represent XOR-/AND-/OR-splits and joins in causal nets. Fig. 4 shows the model inferred considering all students. This model is characterized by a low degree of structure where exams are mainly represented in parallel with each other. This suggests that it is not possible to determine significantly similar trends in students' careers. The only trends are that "C1_1" is typically taken before "A1_2_b" and that "G2_1" is often taken before the end of the third semester.

The model for success students (Fig. 5) looks more structured. Usually, these students took at least "A1_1" and "E1_2" by the end of the first year and "G2_1" and "D1_2" by the third semester. Moreover, "C1_1" is usually taken at least before the end of the second year, sometimes (more precisely, in 30% of the cases) together with "B1_1". Therefore, at least approximately 40% of the exams of the first two years are commonly taken by the end of the second year by this student group. However, we can still observe some exams for which it is not possible to extract a well-defined trend, such as "C1_2_b" and "A1_2_b".

Finally, the model for late-failure students (Fig. 6) looks as much unstructured as the model for all students. We can observe that the model shows strong relations both between "G2_1" and the end of the third semester and between the end of the second year and "G2_1", thus suggesting that a notable percentage of late-failure students were able to take this course on time but, at the same time, another relevant percentage postponed it after the second year.

4.3. Compliance analysis

An analysis of the adherence of students to the manifesto shows an overall fitness of 61.15%. This low value indicates that many students encountered significant challenges in following the manifesto. This result is consistent with the previous analysis. Fig. 8 shows the distribution of fitness for the entire group of students, failure students, and success students. The distribution for the entire group of students
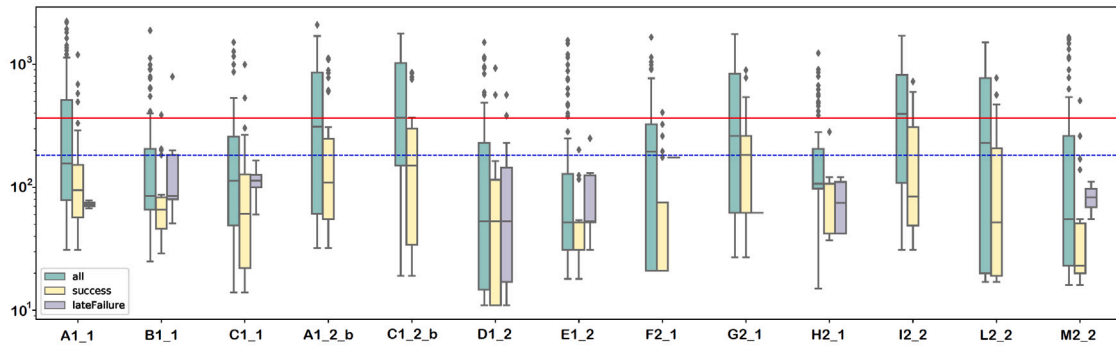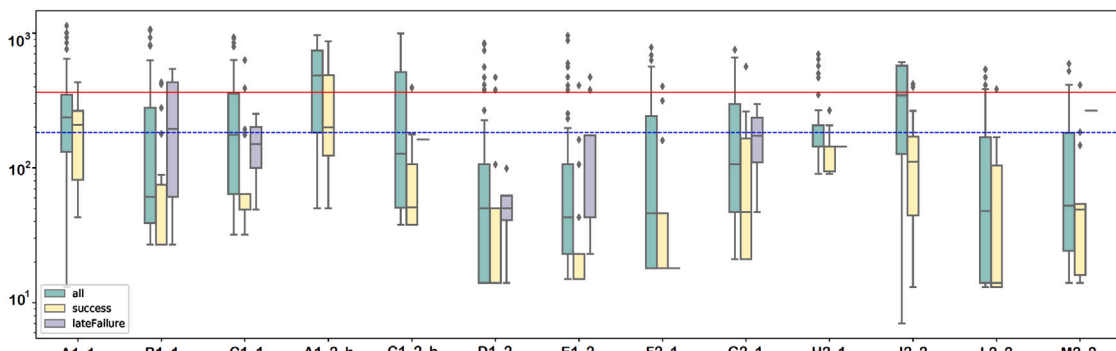
(a) Students enrolled in 2011
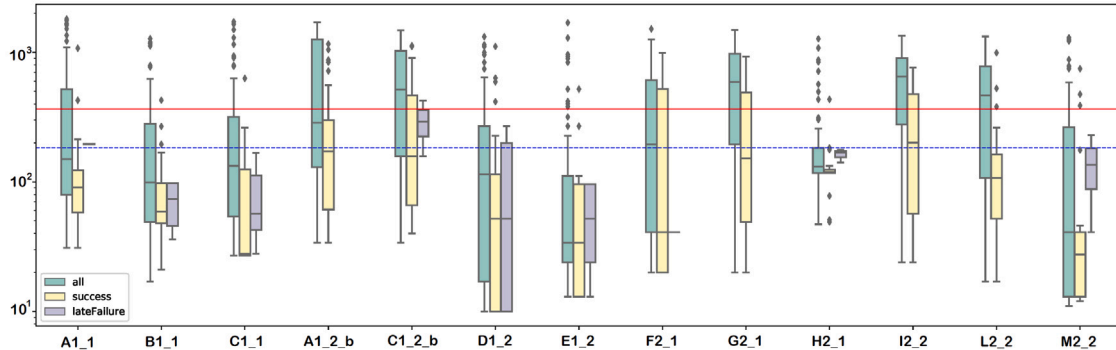
(b) Students enrolled in 2012
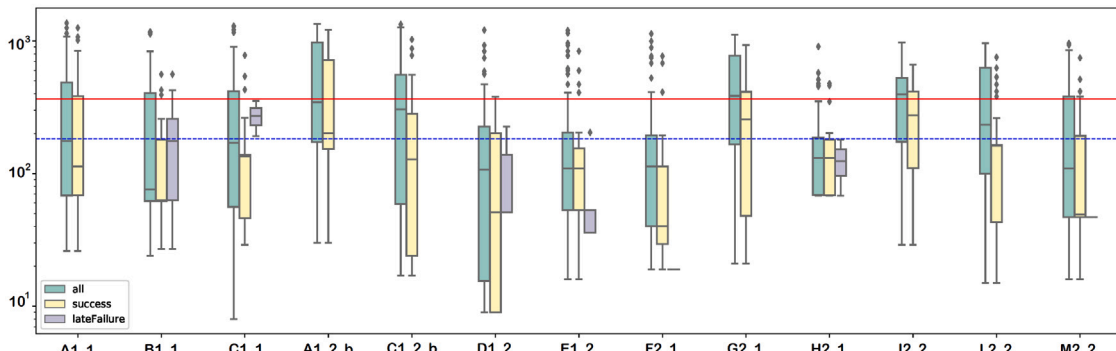
(c) Students enrolled in 2013

(d) Students enrolled in 2014

Fig. 3. Analysis of the time (expressed in days) between the end of a course and the completion of the corresponding exam. The blue and red horizontal lines indicate a six-month and one-year delay from the end of the course, respectively.

is spread between low-medium values, approximately ranging from 0.5 to 0.7, with a median of 0.59. The maximum value is around 0.9, and only a few outliers turned out to be close to 1. The fitness for the failure group mainly ranges between 0.5 and 0.6, with a median

of 0.54 and a maximum value of around 0.69. On the opposite, the distribution of the fitness values from the success group mainly ranges between 0.7 and 0.8, with a median equal to 0.74. The maximum value is 1. The striking difference between the fitness distribution of the

Table 6
Statistics of the event logs of the students.

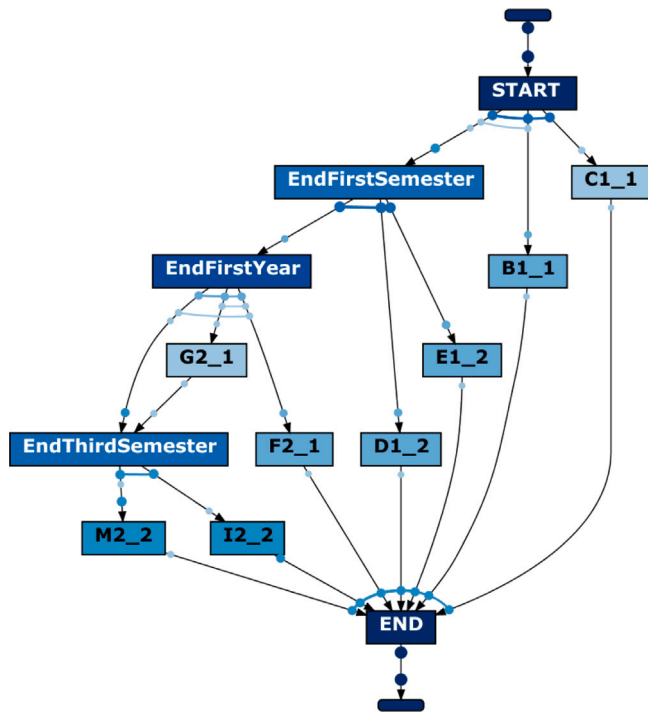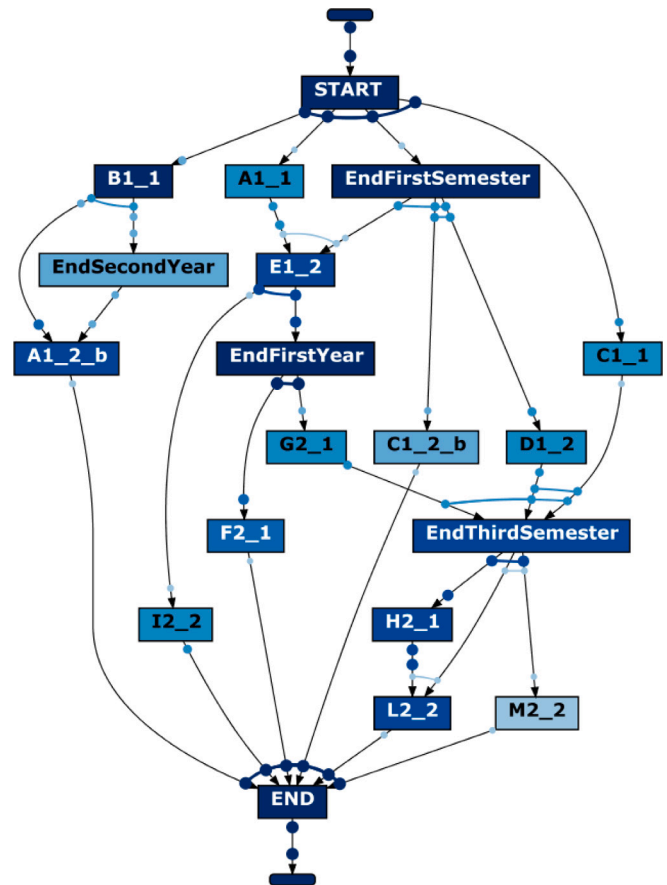| Category	#Students	#Events	Min. Activities	Max. Activities	Mean. Activities
All	510	7462	7	19	15
Success	132	2508	19	19	19
Late failure	54	457	7	14	8

Fig. 4. Students' careers represented as Causal nets.

Fig. 5. Success students' careers represented as Causal nets.

failure and success groups seems to suggest that indeed success students are characterized by much stricter adherence to the manifesto. To determine whether these differences are significant, we ran a statistical test. In particular, we performed the Wilcoxon–Mann–Whitney two-sample rank-sum test (Kruskal, 1957) since the normality assumptions were not met. The test returned a p-value of 2.2^{-16}, which confirms that the differences in fitness are actually statistically significant.

4.4. Lesson learned

The results of the analysis presented in this section allowed us to identify points of weakness and helped the university stakeholders define strategies to improve the performance of the course of study. We discussed the courses exhibiting a low percentage of on-time students with students and faculty members to identify possible issues that led to such performance. With professors, we have defined new teaching programs and activated additional class hours in which optional exercises are discussed. In addition, tutors have been identified among final-year students to support freshmen, especially the ones who have difficulties passing the mandatory first-year exams. An interesting finding of our study is that a somehow more structured career is less likely to fail. Hence, we shared this finding with students and suggested to them the best career to follow. In particular, we presented the results of our analysis to students during the *Welcome days* and explained to them the importance of strictly adhering to a more structured path. It is worth noting that this kind of information was the same information we gave to students every year, explaining to them the importance of following the manifesto. However, we have noticed that seeing results extracted from the actual careers of their peers has greater effectiveness

than an abstract explanation from faculty. Indeed, bringing data to students had a strong impact: we analyzed students' behavior in the following years and we noticed that many of them migrated to the right path. Finally, based on the results of this study and other analyses, we implemented a career monitoring system to identify the weaknesses of the system, evaluate the improvement actions implemented, and support and advise students during their studies. Further studies are needed to evaluate its effectiveness.

5. Related work

Educational Data Mining (EDM) is an emerging discipline that aims to understand and improve students' learning process (Peña-Ayala, 2014; Romero & Ventura, 2007). Dozens of approaches have been proposed, together with empirical studies to evaluate their effectiveness, to address a plethora of different tasks ranging from the construction of social networks describing students' interactions in e-learning activities (Dráždilová et al., 2008) to the profiling of students using course evaluation data (Trandafili et al., 2012) and recommendations on the courses to enroll (Aher & Lobo, 2012). EDM approaches usually apply and adapt classic data mining techniques and concepts, such as clustering, classification, and association rules mining, to educational data. Recent surveys on EDM techniques for the prediction of students'
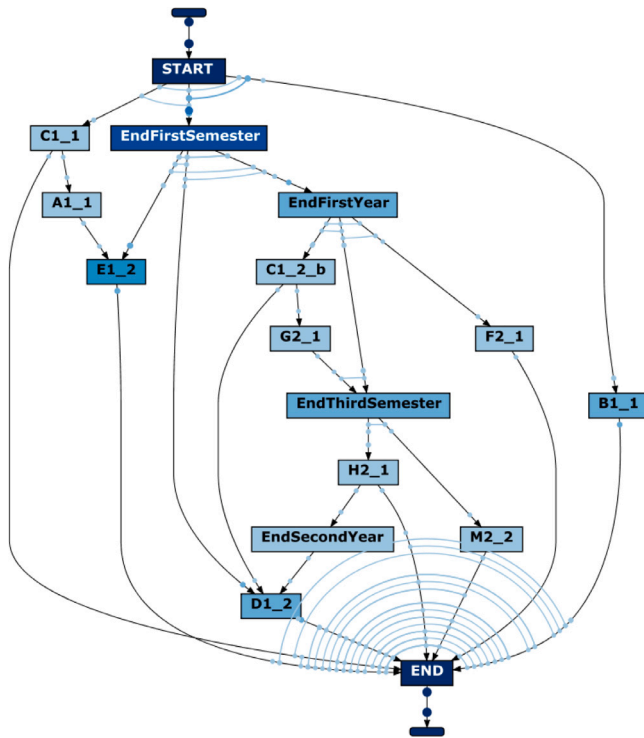
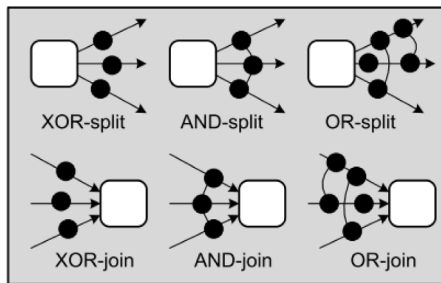**Fig. 6.** Early-failure students' careers represented as Causal nets.

Fig. 7. Notation used for XOR-/AND-/OR-splits and joins.
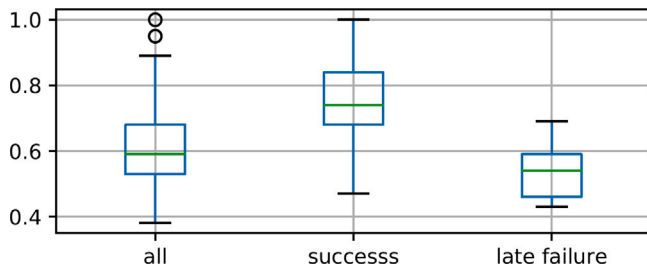

Fig. 8. Distribution of fitness values between success and failure students.

performance can be found in dos Santos Garcia et al. (2019), Xiao et al. (2022).

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 (Dekker et al., 2009; Gowda et al., 2011; Guruler et al., 2010; Herzog, 2005; Lassibille & Navarro Gómez, 2008; Romero et al., 2008). Feature selection, i.e., determining which features are likely to have an impact on students' performance is a crucial phase for those studies.

Several features have been tested in literature, including students' personal characteristics (e.g., gender, age, country), background (e.g., results in high school courses), and academic results (e.g., marks of first-year courses). 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. It is worth noting that even studies centered on the graduation process usually perform a data-oriented analysis, in which students' behaviors are encoded in terms of features without taking into account the underlying structure of the study program.

In this respect, our work is similar to the one of Campagni et al. (2015), who propose to model and analyze students' careers. They introduce the notion of *ideal career* that corresponds to the career of a graduated student who took each exam just after the end of the corresponding course; since some exams might be taken in the same semester, they use an expert to determine the most appropriate orders in which the exams should be taken, thus obtaining a sequence in which each exam is identified by its position. Each student's career is modeled in terms of a sequence of integers, each corresponding to the position that the exam should have had according to the ideal path. Different metrics are used to measure the distance between each student's career and the ideal one (e.g., the Bubblesort distance). These measures are used to infer clusters of students with the aim of exploring possible relations between the distance from the ideal path and academic success. They also exploit sequential pattern mining techniques to infer the most common subsequences of exams. Each element of the sequence corresponds either to the exams taken in the same semester or taken with some delay (measured in terms of semesters). Compared to our approach, the work in Campagni et al. (2015) differs from ours both in terms of the specific goals of the study and the adopted methodology. Campagni et al. (2015) group students taking into account their distance from the ideal path to then analyze differences in performance among the derived groups. On the contrary, our study aims to investigate differences in careers among students classified as successful and not successful according to given performance indicators. Furthermore, their approach does not exploit the potentialities of process-based analysis in modeling students' behaviors. Instead, they apply sequence mining to detect the most frequently followed portions of the actual careers. In contrast, we exploit process formalisms both to model the manifesto of study programs that explicitly accounts for parallelisms, thus allowing us to obtain a more accurate evaluation of the difference between single careers and the ideal path, and to infer start-to-end models representing the overall students' behaviors. It has also to be noticed that their approach is not suitable to find blocks, which is the focus of our work.

The application of process mining techniques to educational data, referred to as *Educational Process Mining* (EPM) (Bogarín et al., 2018), is a subject that has been recently gaining increasing interest. EPM has been applied to deal with different educational problems, such as online learning environments (Bogarín et al., 2014; Deeva & De Weerdt, 2019; Mukala et al., 2015; Real et al., 2021; Vidal et al., 2016), computer-aided online assessments (Bala et al., 2023), computer-supported collaborative learning tools (Bergenthum et al., 2012; Reimann et al., 2009), professional training (Bergenthum et al., 2008; Cairns et al., 2015). However, only a few works investigated the applications of EPM to curriculum mining. Trcka and Pechenizkiy (2009) propose a set of patterns modeling typical constraints of academic curricula and use these patterns to analyze the graduation process (e.g., whether students' behaviors fit those patterns). Our analysis presents some similarities with this approach; we also exploit a reference model modeling the study program to assess the compliance of students' careers. However, we infer models representing students' careers and perform an analysis of students' delays, which are neglected in Trcka and Pechenizkiy (2009). Azeta et al. (2022) apply process mining techniques to analyze event log data generated within educational information systems, with the purpose of understanding

students' behavior during online learning. The work differs from ours in two main ways: (i) it is based on the concept of digital twin for the representation of students' activities and (ii) its focus is on the single course while our work focuses on the entire career.

To the best of our knowledge, only a few works consider the entire students' career. Priyambada et al. (2021) focus on changes in students' learning behaviors over time. For 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. Salazar-Fernandez et al. (2021) propose 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 modeling 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 Hobeck et al. (2023) 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 or late students, and they do not focus on analyzing bottlenecks in the study program. Cameranesi et al. (2017) applies process discovery techniques to curriculum event logs with the purpose of characterizing behaviors of students that performed best/worst in terms of years required to complete the graduation process and final grade. In this work, we shift the focus to classes of students defined according to the indicators defined by the Italian Minister of Education. Moreover, we investigate students' compliance with the manifesto and the students' delays in taking their exams.

6. Conclusions and future work

In this work, we investigated students' careers in a Bachelor program at an Italian university to determine common bottlenecks and potential causes of delays in students' graduations. We applied process mining techniques, in particular process discovery and compliance analysis techniques, to extract and compare the careers of successful and late students. Our analysis allows us to determine common bottlenecks that seem to have an impact on students' graduation time. Moreover, we were able to determine the curriculum path distinguishing successful and late students. The insights gathered by this analysis can be used to support university personnel in delving into factors causing some exams to be a bottleneck, as well as to determine potential improvements in the courses and, eventually, in the overall curricula. Our results provide evidence-based observations that can be used by university stakeholders to provide recommendations to students on how to schedule their exams to avoid incurring delays. In fact, such results allowed us to identify points of weakness and helped the university to define strategies to improve the performance of the course of study. In particular, we have defined new teaching programs and activated additional class hours for optional exercises. In addition, tutors have been identified among final-year students to support freshmen, especially students who have difficulties passing the mandatory first-year exams. An interesting finding is that a somehow more structured career is less likely to fail. Hence, we shared this finding with students to suggest to them the best career to follow. We analyzed students' behavior in the following years and we noticed that many of them migrated to the right path. Finally, based on the results of this study and other analyses, we implemented a career monitoring system to identify the weaknesses of the system, evaluate the improvement actions implemented, support and advise students

during their studies. In future work, we plan to extend the analysis to consider additional elements that can provide us with a better understanding of how and when students prepare themselves for exams. This can be achieved, for example, by keeping into account how many times a student enrolled himself for an exam and analyzing whether the enrollment occurs to the first available window or later. Taking the number of attempts into account can also provide additional insights into the degree of difficulty of a given exam, as well as providing information on whether the students mostly have to retake exams that they fail or are more interested in improving their results. Another interesting research direction is represented by the separate study of the stumbling blocks in first-year versus second-year cohorts. In fact, first-year students often face different and non-academic stumbling blocks (e.g., anxiety as a result of new work schedules and places). In that case, the actual constraints are not programmatic and technical, but rather economic, social, environmental, and psychological.

CRediT authorship contribution statement

Claudia Diamantini: Conceptualization, Methodology, Supervision. **Laura Genga:** Data curation, Methodology, Writing – original draft. **Alex Mircoli:** Writing – original draft , Writing – reviewing & editing. **Domenico Potena:** Conceptualization, Validation, Supervision. **Nicola Zannone:** Methodology, Writing – reviewing & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

References

Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca (2021). Indicatori CdS. https://www.anvur.it/attivita/ava/indicatori-di-monitoraggio-autovalutazione-e-valutazione-periodica/indicatori-cds/.

Aher, S. B., & Lobo, L. (2012). Applicability of data mining algorithms for recommendation system in e-learning. In *Proceedings of international conference on advances in computing, communications and informatics* (pp. 1034–1040). ACM.

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

Azeta, A., Agono, F., Adesola, F., Nwaocha, V., & Tjiraso, S. (2022). A process mining framework for analysing students' behaviours using digital twin. In *Proceedings of international conference on information systems and emerging technologies*.

Bala, S., Revoredo, K., & Mendling, J. (2023). Process mining for analyzing open questions computer-aided examinations. In M. Montali, A. Senderovich, & M. Weidlich (Eds.), *Process mining workshops* (pp. 565–576). Cham: Springer Nature Switzerland.

Bergenthum, R., Desel, J., Harrer, A., & Mauser, S. (2008). Learnflow mining. In *DeLFI 2008: Die 6. E-Learning fachtagung informatik* (pp. 269–280). Gesellschaft für Informatik eV.

Bergenthum, R., Desel, J., Harrer, A., & Mauser, S. (2012). Modeling and mining of learnflows. In *Transactions on petri nets and other models of concurrency V* (pp. 22–50). Springer.

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

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In *Proceedings of international conference on learning analytics and knowledge* (pp. 11–15). ACM.

Cairns, A. H., Gueni, B., Assu, J., Joubert, C., & Khelifa, N. (2015). Analyzing and improving educational process models using process mining techniques. In *Proceedings of international conference on advances in information mining management* (pp. 17–22).

Cameranesi, M., Diamantini, C., Genga, L., & Potena, D. (2017). Students' careers analysis: A process mining approach. In *Proceedings of international conference on web intelligence, mining and semantics* (p. 26). ACM.

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

Deeva, G., & De Weerdt, J. (2019). Understanding automated feedback in learning processes by mining local patterns. In F. Daniel, Q. Z. Sheng, & H. Motahari (Eds.), *Business process management workshops* (pp. 56–68). Cham: Springer International Publishing.

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

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

Dráždilová, P., Martinovic, J., Slaninová, K., & Snášel, V. (2008). Analysis of relations in elearning. In *Proceedings of international conference on web intelligence and intelligent agent technology, vol. 3* (pp. 373–376). IEEE.

Gowda, S., Baker, R., Pardos, Z., & Heffernan, N. (2011). The sum is greater than the parts: Ensembling student knowledge models in ASSISTments. In *Proceedings of KDD workshop on knowledge discovery in educational data*.

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

Herzog, S. (2005). 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.

Hobeck, R., Pufahl, L., & Weber, I. (2023). Process mining on curriculum-based study data: A case study at a German University. In *Process mining workshops* (pp. 577–589). Cham: Springer Nature Switzerland.

Kruskal, W. H. (1957). Historical notes on the Wilcoxon unpaired two-sample test. *Journal of the American Statistical Association*, *52*(279), 356–360.

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

Ministero dell'Università e della Ricerca (2019). Decreto Ministeriale del 7 gennaio 2019 n. 6. Autovalutazione, valutazione, accreditamento iniziale e periodico delle sedi e dei corsi di studio. Allegato E..

Mukala, P., Buijs, J. C., Leemans, M., & van der Aalst, W. (2015). Learning analytics on coursera event data: A process mining approach. In *Proceedings of international symposium on data-driven process discovery and analysis* (pp. 18–32).

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

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

Real, E. M., Pimentel, E. P., & Braga, J. C. (2021). Analysis of learning behavior in a programming course using process mining and sequential pattern mining. In *IEEE frontiers in education conference* (pp. 1–9). IEEE, http://dx.doi.org/10.1109/FIE49875.2021.9637146.

Reimann, P., Frerejean, J., & Thompson, K. (2009). Using process mining to identify models of group decision making in chat data. In *Proceedings of international conference on computer supported collaborative learning* (pp. 98–107). International Society of the Learning Sciences.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135–146.

Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In *Proceedings of international conference on educational data mining* (pp. 8–17). www.educationaldatamining.org.

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

Sanjeev, A. P., & Zytkow, J. M. (1995). Discovering enrollment knowledge in university databases. In *Proceedings of international conference on knowledge discovery and data mining* (pp. 246–251).

Trandafili, E., Allkoçi, A., Kajo, E., & Xhuvani, A. (2012). Discovery and evaluation of student's profiles with machine learning. In *Proceedings of Balkan conference in informatics* (pp. 174–179). ACM.

Trcka, N., & Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. In *Proceedings of international conference on intelligent systems design and applications* (pp. 1114–1119). IEEE.

van der Aalst, W. (2011). Process mining - discovery, conformance and enhancement of business processes. Springer.

van der Aalst, W., Adriansyah, A., & van Dongen, B. (2011). Causal nets: A modeling language tailored towards process discovery. In *Proceedings of international conference on concurrency theory* (pp. 28–42). Springer.

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.

van der Aalst, W., van Dongen, B., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. J. (2003). Workflow mining: A survey of issues and approaches. *Data & Knowledge Engineering*, *47*(2), 237–267.

van Dongen, B., De Medeiros, A. A., & Wen, L. (2009). Process mining: Overview and outlook of petri net discovery algorithms. In *Transactions on petri nets and other models of concurrency II* (pp. 225–242). Springer.

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

Weijters, A., van der Aalst, W., & De Medeiros, A. A. (2006). *Process mining with the Heuristics Miner-algorithm*: *Technical report 166*, Technische Universiteit Eindhoven.

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