# Modelling Computer Engineering Student Trajectories with Process Mining

Pablo Martinez[1], Oscar Montañes[1], Juan Manuel Serralta[1] and Libertad Tansini[1]

[1]Computer Science Institute (Inco), Faculty of Engineering, Universidad de la República, Montevideo, Uruguay

**Abstract**

This work presents the analysis of student learning trajectories in their Computer Engineering studies. The analysis focuses on modelling characteristics that have impact on dropout rates. Hence students trajectories of *dropout* students and *graduated* students are analyzed and compared using Process Mining tools. Specific course that are considered "difficult" and prevent academic progress are identified, also the last courses with which students dropout or graduate. Some of the courses identified as "bottle necks" are *Sistemas Operativos*, *Redes de Computadoras* and *Arquitectura de Computadores*. And some of the courses that are left for last are *Física 1* and *Métodos Numéricos*. The results show that dropout students manly finish after the first year and that they choose courses related to Programming. The models adequately describe student trajectories, with the usual metrics for Process Mining of fitness over 97% on the trajectories of the log, and do not overfit the training data set showing high generalization.

**Keywords**

Student Learning Trajectories, Computer Engineering Degree, Process Mining, Modelling

## 1. Introduction

The motivation for this project is the need of the Education Unit of the Engineering Faculty (Unidad de Enseñanza de Facultad de Ingeniería, de la Universidad de la República, Uruguay (UEFI)) to explain the causes for dropout of students enrolled in the different degrees, based on the available information in their Information System. This work is part of a larger project [1] where the reports required by UEFI were automatized through the implementation of a first version of a Data Warehouse; an in depth descriptive analysis of the variables that may have greater incidence in student dropout was carried out, exploring new data sources (such as Continuous Household Survey (ECH), primary shool scholarship from ANEP (Administración Nacional de Educación Pública, https://www.anep.edu.uy/) and geo-referencing of the addresses of the students); and finally the modelling of students trajectories and behaviour along their studies in their Computer Engineering studies was made, using machine learning and process mining techniques to analyze possible social and curricular reasons that explain dropout.

This paper explains with more detail the last aspect of the project, that is, the analysis of trajectories of Computer Engineering students using the process mining tool ProM Tools

*IV LATIN AMERICAN CONFERENCE ON LEARNING ANALYTICS*

p.martinezben@gmail.com (P. Martinez[1]); omontanes@gmail.com (O. Montañes[1]); juanserralta1@gmail.com (J. M. Serralta[1]); libertad@fing.edu.uy (L. Tansini[1])

0000-0002-1212-3871 (P. Martinez[1]); 0000-0002-7893-823X (O. Montañes[1]); 0000-0001-5564-7352 (J. M. Serralta[1]); 0000-0001-6017-0114 (L. Tansini[1])

[2] to model an analyze the data in the “Administrative Information System” (AIS) of the University, also called *“Bedelía”*, “University administrator office”, “Admissions office”, “Office of the Registrar”, etc.

62% of the students dropout from their Computer Engineering studies. The goal of this work is to provide analytical tools for decision makers to understand the reasons why students leave. To achieve this goal, *dropout* and *graduated* students trajectories were modelled and analyzed in order to identify courses that hinder students from advancing or graduating, and courses with which students dropout or graduate.

The results show **general behavioural patterns of the students** during their studies, identifying hard courses to pass or “bottle neck” courses such as *Sistemas Operativos*, *Redes de Computadoras* and *Arquitectura de Computadores.* Also courses that are left for last, are found, such as *Física 1* and *Métodos Numéricos.* The results show that dropout students manly leave without finishing the first year and those who advance more **manage to pass courses related to Programming**. The models adequately describe the trajectories with a fitness above 97% on the log traces, they do not overfit the provided training logs and have high generalization.

The following sections describe related works, methodology, results, and finally conclusions and future work.

## 2. Works related to Learning Analytics and Process Mining

In Uruguay, tens of thousands of students annually join the education system with three or four years, starting in this way their “educational life”. Many of them will finish primary school, less will finish high school and only a few will achieve university degrees. Information is naturally produced and registered by students and teacher (grades, daily assistance, evaluation, etc.) in administrative or pedagogical information systems, recording students educational ”traces”[3].

For some decades now, there have been several initiatives advocating for the application of technologies such as Data Mining, Artificial Intelligence, Big Data, among others to the analysis of a variety of social problems, including education [4], in what has been called *“Learning Analytics”* (LA) [5]. Privacy of information and regard for personal information is only one of the many challenges of LA [6]. Several authors have used process mining or *“Educational process mining”* (EPM) to model different behavioural aspects of students trajectories, for example in the analysis of LMS (Learning Management Systems) logs [7], a recent survey of different application areas and tools can be found in [8].

## 3. Methodology

ProM Tools is used as the process mining tool to model students trajectories from the data in the “Administrative Information System” (AIS), with the aim to identify “bottle neck” courses, and given the flexibility to choose courses, find courses that are left for last and specific ordering patterns among curses for the different groups of students.

**Table 1**
Necessary information for Process Mining.

| Column | Type | Description |
|---|---|---|
| id_number | INT | Id number |
| course_id | VARCHAR | Name of the course |
| date_start | DATE | Date of first register |
| date_end | DATE | Date approval |
| status | VARCHAR | State of the student<br>(DROPOUT, ATTENDING, GRADUATED) |

### 3.1. The AIS Data Base

In several meetings with the Education Unit of the Engineering Faculty (UEFI), having previously signed a confidentiality agreement, it was possible to have access to the information in the AIS Data Base. It is an SQLite base with various tables that describe the students activities, the most important ones are described in this section.

**Table Activ2:** contains all activities of the students in the different degrees they are enrolled in. The activities are ordered by date and include course inscriptions and exams, with their corresponding result and grade.

**Table Estudiante:** this table contains Student entities with all its personal information, such as birth date, gender, were they pursued high school studies and contact information.

**Table Estudiante-Carrera:** this table registers the students dates of enrollments and graduation in different degrees. There are over ten degrees.

**Table Asignaturas:** contains the codes of the courses in the Engineering Faculty and their relationship with the different degrees. There may be different evaluation methods and credits per course in the different degrees.

With this set of tables it is possible to determine the scholarship of the students to load the necessary information in the process mining tool. That is, to determine each of the course in which the students have enrolled, have passed or failed, the number of each event and the dates.

### 3.2. Extraction, Transformation and Loading (ETL)

It was necessary to transform, clean and unify the data in order to load it into ProM Tools since the structure of the before mentioned tables does not allow to obtain in a simple manner the scholarships of the students. Table 1 shows the right format describing the courses passed by the students, i.e. their scholarships.

The following transformations, cleaning and unification operations were performed:

- **Elimination of trajectories with transferred courses**: course credits have been transferred from other schools or degrees, these courses are relatively few and distort the final models.
- **Elimination of optional courses**: they represent a small portion of the total number of courses taken by the students and are to heterogeneous to model adequately in this stage.

- **Mandatory courses**: the suggested courses by AIS [9] is used as the set of possible courses for the analysis.
- **Unification of courses**: many courses undergo changes along the years, having different versions, with different names, codes and content. For example *Cálculo 1* was named *Calculo Dif. e Integral en una Variable* from the year 2017 and on, with a slight change in the content. For this analysis, those differences are not relevant, hence, for this study all related courses were grouped together.

**Table 2**
Conversion from CSV to XES format.

| Column | Description |
|---|---|
| *Case_id* | maps to id_number of the student |
| *Event_id* | maps to course_id |
| *Date_finish* | maps to approval date of the course or date_end |

The data is extracted in CSV format [10], transformed and loaded to ProM, where it is processed. To apply the different process mining algorithms provided by the process mining tool it is necessary to transform the raw data (CSV) to XES [11], which is the format required by ProM. This allows to represent events based on a XML format. The conversion is made using a plug-in provided by ProM, for which it is necessary to have the information shown in Table 2 and which is directly associated to the data provided in the log, see Table 2.

After several meetings with UEFI and the Academic Program Director of the Computer Engineering Degree it was decided to divide the students in three groups to be able to model them adequately: *dropout*, *advanced* and *graduated*. This strategy allows to find specific models for each group instead of pursuing more generalised models for all students together. It is specially relevant to procure analytical tools for the dropout group of students that represents around 62% of the students. To this end, models are built and analyzed for the *dropout* and *graduated* groups of students.

### 3.3. Methodology for the Modelling and Analysis

In the first place the dependency model for the provided logs are obtained with the plug-in Interactive Data-Aware Heuristic Miner [12], which allows the application of different algorithms for modelling, besides offering the final model with different process representation alternatives (Dependency nets, Petri nets, Causal nets, etc.). The generated dependency graph with the Flexible Heuristic Miner [13] as discovery algorithm, gives a first approximation to the characteristics of the underlying model in the log.

Then, with the objective to model the trajectories of the students, several algorithm were analyzed that allow to explore and discover underlying models in the provided logs. In this study, an inductive algorithm was used [14], this was done through the plug-in “Mine Petri net with Inductive Miner”, producing a Petri-Net as a result. The choice was based on previous works in the same topic, such as “Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs)” [14]. The default configuration was used, unless stated.

Finally, for performance tests on the models werenmade with the plug-ins “Replay a log on Petri Net for Conformance/Performance analysis” and “Measure precision/generalization”.

### 3.4. Metrics

The following metrics were used over the models to evaluate conformity and performance:

- **Fitness**: percentage of traces that are recognized by the model.
- **Deviations**: set of activities that present deviations with respect to the model.
- **Average Throughput**: average duration of the traces.
- **Bottle Neck Detection**: courses that require more time to pass.
- **Precision**: how precisely the model represents the observed process.
- **Generalization**: how well the model reproduces future behaviour, confidence in the precision.

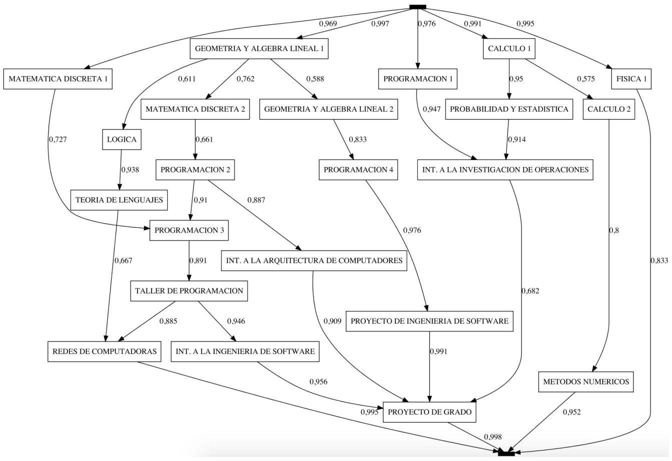

**Figure 1:** Dependency graph of the courses of the *graduated* group of students.

## 4. Results

There is information available of 3477 students between the years 1997 and 2019. In total they produce a log with 33396 events. In the following sections the models for the *graduated* and the *dropout* groups of students are presented and analyzed.

### 4.1. Graduated students

This group is made of 684 students (approximately 20%), with a total of 16416 events (approximately 49%), since they are the ones with most registered activity. An initial inspection shows that 45% of the entrances correspond to passing the basic math course *Geometría y Álgebra Lineal 1* as the first course they pass. Most of the students, around 67% of them, finish with *Proyecto de Grado* which is the final Thesis. It is also interesting to mention that *Redes de Computadoras* y *Métodos Numéricos*, appear second and third as the last courses approved to finish the degree.

The first model shows the dependency graph and presents the dependency relations between courses. As can be seen in Figure 1, in the graph obtained from the log, the rectangles represent the courses and the values on the arrows represent the confidence regarding the dependency relations (the higher value, the more confidence). With this model it is possible to identify common trajectories for students who graduate.

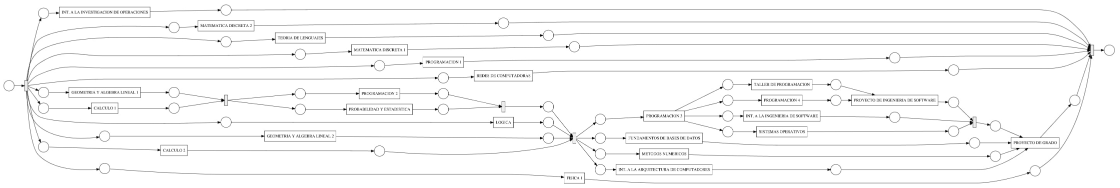
INT. A LA INVESTIGACION DE OPERACIONES
MATEMATICA DISCRETA 2
TEORIA DE LENGUAJES
MATEMATICA DISCRETA 1
PROGRAMACION 1
REDES DE COMPUTADORAS
GEOMETRIA Y ALGEBRA LINEAL 1
CALCULO 1
PROGRAMACION 2
PROBABILIDAD Y ESTADISTICA
LOGICA
GEOMETRIA Y ALGEBRA LINEAL 2
CALCULO 2
FISICA 1
TALLER DE PROGRAMACION
PROGRAMACION 4
PROYECTO DE INGENIERIA DE SOFTWARE
PROGRAMACION 3
INT. A LA INGENIERIA DE SOFTWARE
SISTEMAS OPERATIVOS
FUNDAMENTOS DE BASES DE DATOS
METODOS NUMERICOS
INT. A LA ARQUITECTURA DE COMPUTADORES
PROYECTO DE GRADO

**Figure 2:** Petri net model for the *graduated* group of students.

The following analysis is based on Petri nets [15], with the aim to identify relevant information within the process. The analysis focuses on the metrics: Fitness, Deviations, Average Throughput, Bottle Neck Detection, Precision and Generalization. Figure 2 shows the Petri net obtained with the plug-in "Mine Petri net with Inductive Miner" on the log.

**Table 3**
Average approval times for the *graduated* group of students.

| Curses | Average approval time (years) |
|---|---|
| *Métodos Numéricos* | 2.5 |
| *Introducción a la Ingeniería de Software* | 2.3 |
| *Proyecto de Grado* | 2.1 |
| *Fundamentos de Bases de Datos* | 1.9 |
| *Sistemas Operativos* | 1.9 |
| *Proyecto de Ingeniería de Software* | 1.3 |
| *Programación 2* | 1.2 |
| *Probabilidad y Estadística* | 1.1 |
| *Arquitectura de Computadores* | 1.0 |
| *Taller de Programación* | 1.0 |
| *Programación 4* | 0.8 |
| *Programación 3* | 0.6 |

The conformity of the model for the provided log shows that the model reproduces 97% of the traces, where 375 of the 648 analyzed traces align perfectly with the model. Also 9 of the 24 courses are aligned with the model and the log, and the rest present some deviations since some transitions were detected only in the model and not in the log. In this cases the deviation is less 5% of the traces. The performance analysis shows that the execution time of the traces (throughput) in average for all students is 8 years, the minimum is 4.5 years and the maximum is 20 years. The waiting time analysis shows there are 12 courses with high or very high time of approval, as ca bee seen in Table 3.

The waiting time analysis for the courses shows relevant information regarding bottle necks in the model. Revealing the courses *Sistemas Operativos*, *Fundamentos de Bases de Datos*, *Proyecto de Grado*, *Introducción a la Ingeniería de Software* and *Métodos Numéricos* exhibit approval times four times above expected, considering they should be passed in one semester.

Finally, the conformity metrics of the model give a Precision of 0.30126 and a Generalization of 0.8949, indicating the models do not not overfit the training data and that they are capable of reproducing behaviour not present in the original log, being flexible enough to model new traces.

## 4.2. Dropout students

For this analysis there were 2158 students (approximately 62% of the total), with in total 8266 events (approximately 25%).

A primary inspection of the log of the *dropout* group of students, shows that 69% of them have as the last course one of the courses of the first year: *Cálculo 1*, *Geometría y Álgebra Lineal 1*, *Física 1*, *Programación 1* or *Matemática Discreta 1*. This fact alone proves most of them do not advance further than the first year.

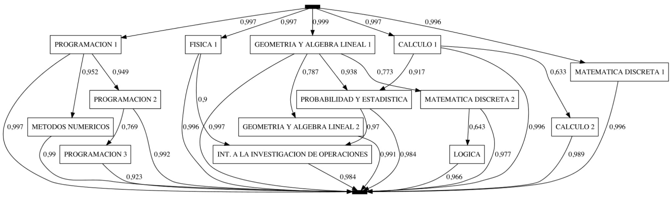

**Figure 3:** Dependency graph of the courses of the *dropout* group of students.

The dependency model that gives a first approximation to the characteristics of the underlying model in the log, was made with "Mine Petri net with Inductive Miner" on the log and with the default parameters, displayed only fist year courses, because most students dropout the first year.

Then the minimum frequency for a course to be considered by the algorithm is lowered, in order to have more courses visible in the dependency model in Figure 3. The model shows that those students that advance the most, manage to pass programming courses, specifically *Programación 1* to *Programación 3*.

Figure 4 shows the Petri net generated with the inductive miner on the log for the *dropout* group of students, to perform conformity and performance analysis.

The resulting model allows to represent 97% of the traces in the log. The courses *Cálculo 1*, *Geometría y Álgebra Lineal 1*, *Programación 1* and *Matemática Discreta 1* are the courses with highest frequency. Among these courses, only *Geometría y Álgebra Lineal 1* and *Matemática Discreta 1* show deviations from the model, less than 33% and 0.1% respectively.

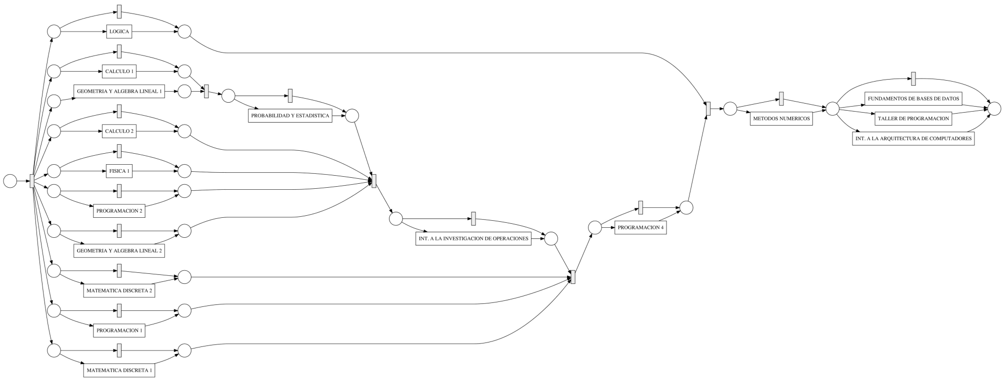
LOGICA
CALCULO 1
GEOMETRIA Y ALGEBRA LINEAL 1
PROBABILIDAD Y ESTADISTICA
CALCULO 2
FISICA 1
PROGRAMACION 2
GEOMETRIA Y ALGEBRA LINEAL 2
MATEMATICA DISCRETA 2
PROGRAMACION 1
MATEMATICA DISCRETA 1
INT. A LA INVESTIGACION DE OPERACIONES
PROGRAMACION 4
METODOS NUMERICOS
FUNDAMENTOS DE BASES DE DATOS
TALLER DE PROGRAMACION
INT. A LA ARQUITECTURA DE COMPUTADORES

**Figure 4:** Petri net model for the *dropout* group of students.

The performance analysis of the model shows that in average, students dropout in 20.16 months or approximately one and a half year. The courses with high approval times are: *"Probabilidad y Estadística*, *Introducción a la Investigación de Operaciones*, *Programación 4*, *Métodos Numéricos*, *Fundamentos de Bases de Datos* and *Taller de Programación*, see Table 4. It is worth to mention that very few of the dropout students reach the more advanced courses in the list, hence more information is needed to completely understand the results.

**Table 4**
Average approval times for the *dropout* group of students.

| Curse | Average approval times (years) |
|---|---|
| *Probabilidad y Estadística* | 1.5 |
| *Taller de Programación* | 1.6 |
| *Métodos Numéricos* | 2 |
| *Fundamentos de Bases de Datos* | 2.4 |
| *Programación 4* | 2.9 |
| *Introducción a la Investigación de Operaciones* | 3 |

The models show Precision of 0.41671 and Generalization of 0.99775, which are superior to the graduated students. The models do not overfit the log and have high generalization capabilities.

## 5. Conclusions and future work

**ProM turned out to be versatile and flexible**, allowing general analysis of the data, the generation of different models and a variety of metrics over them. Nevertheless, it is necessary to have deep understanding of the information of the students, to perform an **adequate prepossessing and cleaning, and finally an adaptation to the required format** as entry to ProM for the results to be useful.

The models give **insight into the learning trajectories or behaviour of the students** in their Computer Engineering studies, enabling the identification off **“bottle neck courses”** or hard courses to pass, as well as courses that do not hinder students from advancing, like *Física 1* and *Métodos Numéricos*. Some of the “bottle neck courses” are *Sistemas Operativos*, *Redes de Computadoras* and *Arquitectura de Computadores*.

Considering the *dropout* group of students, it was possible to verify, both with statistical methods and process mining models, that most of them dropout the first year and that those that advance the most **choose courses related to programming**.

The models adequately describe the student information with fitness over 97% on the log traces, they do not overfit to the log and allow to recognize other traces than those in the training log with high precision and generalization.

For future work it desirable to **include updated information** to the models, since for organizational restrictions it was only possible to work with data until 2019. It is possible to model other aspects of the learning trajectories by including **all of the courses**, such as the optional and then transferred ones that were excluded in this work.

ProM has a series of plug-ins that have not been tested and could be studied for the purpose of modeling student behaviour.

It is of great interest to consider **other information** than that in the AIS Data Base to be added to the models, such as gender, age, work information, income, primary and high school information.

Finally, we aim at exploring the utility of the models segmented by semesters or years to obtain more details.

## References

[1] P. Martínez, O. Montañés, J. Serralta, Modelado de trayectorias académicas de estudiantes universitarios mediante técnicas de analítica de aprendizaje, Tesis de grado. Universidad de la República (Uruguay). https://hdl.handle.net/20.500.12008/28848 (2021).

[2] Prom tools home page, 2020. URL: http://www.promtools.org/doku.php?id=start, (Accessed on 03/12/2020).

[3] Del papel a la nube: Cómo guiar la transformación digital de los sistemas de información y gestión educativa (siged), 2019. URL: https://publications.iadb.org/es/del-papel-la-nube-como-guiar-la-transformacion-digital-de-los-sistemas, -Banco Interamericano de Desarrollo (Accessed on 10/10/2019).

[4] I. Jara, J. Ochoa, Usos y efectos de la inteligencia artificial en educación, Sector Social división educación. Documento para discusión número IDB-DP-00-776. BID. doi: http://dx.doi. org/10.18235/000238 0 (2020).

[5] P. Siemens, George y Long, Penetrating the fog: Analytics in learning and education., EDUCAUSE Review 46 (2011) 30.

[6] A. Pardo, G. Siemens, Ethical and privacy principles for learning analytics, British Journal of Educational Technology 45 (2014) 438–450.

[7] C. Romero, R. Cerezo, A. Bogarín, M. Sánchez-Santillán, EDUCATIONAL PROCESS MINING: Applications in Edu. Research, 2016, pp. 1–28. doi:10.1002/9781118998205.ch1.

[8] A. Bogarín, R. Cerezo, C. Romero, A survey on educational process mining,

WIREs Data Mining and Knowledge Discovery 8 (2018) e1230. URL: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1230. doi:https://doi.org/10.1002/widm.1230. arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1230.

[9] Trayectoria sugerida para la carrera en ingeniería en computación, plan 97, 2020. URL: https://www.fing.edu.uy/carreras/grado/computacion/implementacion/archivos/TrayectoriaSugerida.pdf, (Accessed on 15/12/2020).

[10] Csv, comma separated values file, 2020. URL: https://tools.ietf.org/html/rfc4180#section-2, (Accessed on 15/12/2020).

[11] Xes, extensible event stream, 2020. URL: http://xes-standard.org/, (Accessed on 15/12/2020).

[12] F. Mannhardt, M. de Leoni, H. A. Reijers, Heuristic mining revamped: An interactive, data-aware, and conformance-aware miner., in: BPM (Demos), 2017.

[13] A. J. M. M. Weijters, J. T. S. Ribeiro, Flexible heuristics miner (fhm), 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM) (2011) 310–317.

[14] A. Bogarín, R. Cerezo, C. Romero, Discovering learning processes using inductive miner: A case study with learning management systems (lmss) (2018).

[15] T. Murata, Petri nets: Properties, analysis and applications, Proceedings of the IEEE 77 (1989) 541–580. doi:10.1109/5.24143.