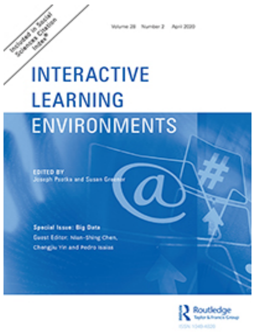


Interactive Learning Environments

ISSN: 1049-4820 (Print) 1744-5191 (Online) Journal homepage: https://www.tandfonline.com/loi/nile20

Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay

Luiz Antonio Macarini, Henrique Lemos dos Santos, Cristian Cechinel, Xavier Ochoa, Virgínia Rodés, Alén Pérez Casas, Pedro Pablo Lucas, Ricardo Maya, Guillermo Ettlin Alonso & Patricia Díaz

To cite this article: Luiz Antonio Macarini, Henrique Lemos dos Santos, Cristian Cechinel, Xavier Ochoa, Virgínia Rodés, Alén Pérez Casas, Pedro Pablo Lucas, Ricardo Maya, Guillermo Ettlin Alonso & Patricia Díaz (2020) Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay, Interactive Learning Environments, 28:2, 166-190, DOI: 10.1080/10494820.2019.1636082

To link to this article: https://doi.org/10.1080/10494820.2019.1636082

Published online: 08 Jul 2019.

INTERACTIVE LEARNING ENVIRONMENTS
2020, VOL. 28, NO. 2, 166–190
https://doi.org/10.1080/10494820.2019.1636082

Towards the implementation of a countrywide K-12 learning analytics initiative in Uruguay

Luiz Antonio Macarini [a], Henrique Lemos dos Santos [b], Cristian Cechinel [a], Xavier Ochoa [c], Virgínia Rodés [d], Alén Pérez Casas [d], Pedro Pablo Lucas[e], Ricardo Maya [e], Guillermo Ettlin Alonso[f] and Patricia Díaz[f]

[a]Coordenadoria Especial Interdisciplinar de Tecnologias da Informação e Comunicação, Universidade Federal de Santa Catarina, Araranguá, Brazil; [b]Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil; [c]New York University Steinhardt School of Culture Education and Human Development, New York, NY, USA; [d]Programa de Entornos Virtuales de Aprendizaje, Comisión Sectorial de Enseñanza, Universidad de la Republica, Montevideo, Uruguay; [e]Escuela Superior Politecnica del Litoral, Guayaquil, Ecuador; [f]Departamento de Tecnologías Digitales, Consejo de Formación en Educación, Administracion Nacional de Educacion Publica, Montevideo, Uruguay

ABSTRACT

The present work describes the challenges faced during the development of a countrywide Learning Analytics study and tool focused on tracking and understanding the trajectories of Uruguayan students during their first three years of secondary education. Due to the large scale of the project, which covers an entire national educational system, several challenges and constraints were faced during its conception and development. Examples of key challenges were: to understand the different nuances of the educational system, to satisfy ethical and legal requirements without narrowing down the scope and potential novelty of the initiative and to deal with integration and inconsistencies in the databases. This paper presents the design decisions and solutions found to address or mitigate the problems found as a contribution to facilitate similar large-scale projects. Three main experiments using data mining were conducted and their results are also described, pointing out the feasibility of finding meaningful patterns that can be used by educational authorities to improve their decision making and foster public educational policies. Finally, the paper describes the use of the data and the findings of the study to create an interactive visual tool to explore the relationship between student variables, performance and persistence in the system.

ARTICLE HISTORY
Received 30 October 2018
Accepted 23 April 2019

KEYWORDS
Academic trajectory; dashboard; educational data mining; learning analytics; visualization; secondary education

1. Introduction

Modern educational systems are engineered to store students' data that can be analyzed using Educational Data Mining (EDM) and Learning Analytics (LA) techniques in order to improve educational settings in different contexts (Alom & Courtney, 2018; Sin & Muthu, 2015). Current literature on Learning Analytics cover a wide range of studies that vary in several dimensions. For example, from the techniques employed (data mining, visualization, social network analysis, statistics), to the source of the data collected (Learning Management Systems, surveys, sensors), the stakeholders involved (students, professors, administrators), and the educational level to which the systems/experiments are directed, just to mention a few. Even though LA is a well-known recognized emergent field, there are still few works manipulating countrywide datasets, e.g. Alom and Courtney (2018), Hutt,

CONTACT Luiz Antonio Macarini ✉ luiz.buschetto@posgrad.ufsc.br

Gardener, Kamentz, Duckworth, and D'Mello (2018), and Howard and Mozejko (2013). This probably has to do with the inherent difficulties and challenges involving the use of large datasets (Agrawal et al., 2011) such as, for instance, inconsistency, incompleteness, heterogeneity and data security (Hlosta, Zdrahal, & Zendulka, 2017; Kouzes, Anderson, Elbert, Gorton, & Gracio, 2009). It may be also related to the very existence (or non-existence) of countrywide educational data that is already organized, structured, prepared and available to be integrated and consumed by LA systems. Issues related to the complex landscape of ethics, privacy and political policies also play an important role here, and are all interconnected with other aspects such as trust, accountability and transparency that need to be considered during the design of LA systems in such a large scale (Pardo & Siemens, 2014). Not to mention the challenge of defining a clear methodological approach for the experiments: LA spanned a myriad of expertise and analytical techniques, but still there is no precise recipe towards good practices on LA. Moreover, the multidisciplinary character of this field can lead to a infinity of fields trying to contribute in some aspects, but also potentially causing some philosophical issues on the models of quality and the quality assessment itself (Gibson & Lang, 2018). When considering a countrywide LA project, methodological choices are obviously also restricted by the constraints previously mentioned, and that directly interfere in the way researchers and developers are able to access and manipulate the data, as well as the way results could be presented.

The Uruguayan Educational System is characterized by universal and free coverage at the primary level. However, over the last decades, it has experienced important problems associated with grade retention and students dropout. At the same time, the system has faced a non-enrollment increase. These problems create serious difficulties that threaten the permanence of a sizable proportion of students in the system (INEEd, 2017).

According to Manyika et al. (2011), *Big Data* can be described as the *datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze*. It is an intentionally subjective concept since the technology is always evolving and the size (measured in bytes, for example) is relative to its epoch. Also, the definition is different for each sector in which the data is used (Manyika et al., 2011). According to Siemens and Long (2011), learning analytics can "*penetrate the fog*", bringing improvements on the quality and value of the learning experience by exploring big data in the educational context. The educational data used throughout this project can be seen as an example of Educational Big Data (EBD) since it fulfills at least four of the five Vs of big data (Storey & Song, 2017): *Volume* – ranging from 2015 to 2017, there are more than 400 thousand distinct students in the database and also more than 700 thousand enrollments data, not to mention the assessment data which accounts for over 2 million records, this volume of data required ad hoc software which will be named later; *Velocity* – recently, at each year around 250 thousand of new enrollments are added to the database, and the assessment data grows at a faster pace since there are several evaluation meetings during a year; *Veracity* – the databases are maintained by the National Administration for Public Education which guarantees a reasonable level of confidence, although several data are still inserted by humans and thus susceptible to mistakes which require concerns about validation and data cleaning; *Value* – the actionable knowledge that can be extracted from the databases can cause major changes. It can reveal, for instance, where the government funds should be directed to or, in a teaching level, which students demonstrate a failing trend already during the first months. Regarding the missing V (*Variety*), the data available also presents some level of variety, despite being structured data, whereas at each evaluation meeting the teachers can describe the student's performance in a free text form. These free text data are not currently being analyzed in this project scope, but still can contain useful information.

Parallel to more sociological approaches, this work is the first step into understanding the problems in the Uruguayan Educational System through a more quantitative approach based on Learning Analytics. The focus of this paper is to present the most important challenges faced during the conduction of a countrywide LA project focused on developing a system to track the academic trajectory of secondary students of Uruguay (K-12 data). Together with these challenges, it will also be presented the initial results obtained from the mining process and how these results were used by the research team inside the system under development. We hope the insights brought to light here help the Learning Analytics community on understanding how to define guidelines and best practices towards the employment of educational data mining and data visualization techniques over a countrywide educational big data.

Despite being primarily an exploratory study, all the efforts put on this project helped on building two main research questions, namely:

- RQ1 – Which are the main challenges in order to implement a countrywide Learning Analytics initiative?
- RQ2 – How feasible is to uncover useful educational metrics from a countrywide database?

The present paper extends previous work (Macarini et al., 2019) in a number of ways. First, we included a related work section that presents similar works in the field of big data and Learning Analytics, together with an analysis of the main characteristics of these existing initiatives. Second, we also present here a better contextualization of the Uruguayan public educational system, its most recent changes and challenges, and a general statistical description of the gathered data. Third, we included an in-depth description of the data mining experiments performed so far, the main results unveiled and the educational metrics encountered. Fourth, the present paper also considers two research questions that are discussed in a new section. At last, here, we present a sneak peak of the visualization tool developed to provide to the governmental authorities interactive ways to visualize and analyze the countrywide K-12 educational data.

The remainder of this paper is organized as follows. Section 2 presents some Learning Analytics initiatives around the globe that deal with Big Data, mostly focusing on those works covering countrywide or governmental data. Section 3 contextualizes the educational system of Uruguay and Section 4 presents information about the database used. Section 5 describes the methodology followed. In Section 6, the three most fruitful experiments and their results are showed. Section 7 discusses the main challenges of the present initiative, answers the proposed research questions and depicts an initial view of the prototype under development. Finally, Section 8 presents the final remarks and future work.

2. Related work

Although there are plenty of works dealing with relatively large educational datasets, just few of them cover a representative portion of an entire country population. Usually these works are focused on a given university, on a specific region or on heterogeneous communities like social networks (Araque, Roldán, & Salguero, 2009; Wu, Hsiao, & Nian, 2018; Xu & Jaggars, 2011). A recent survey about learning analytics initiatives in Latin America (dos Santos, Cechinel, Nunes, & Ochoa, 2017) has shown that most of the local efforts are concentrated on experimentation or on theoretical analysis only. A conclusion that can be drawn from that is twofold: there are challenges not only in accessing these kind of data but also in providing useful applications to the stakeholders as most of the well-known online learning systems offers limited support regarding learning analytics insights (Ali, Asadi, Gaevi, Jovanovi, & Hatala, 2013). In the next paragraphs some recent works related to large educational datasets – considering both quantity and coverage – will be briefly discussed.

In Bezerra, Scholz, Adeodato, Lucas, and Ataide (2016) the authors analyzed the school dropout during the last year of the public elementary school in the state of Pernambuco, Brazil. They applied data mining algorithms over the educational census data (almost 2.7 million registrations), ranging from 2011 to 2012, in an attempt to profile students subjected to dropout. Results showed that factors like age, schedule of the classes and geographical location of the schools strongly influences on the dropout trend. Calixto, Segundo, and de Gusmão (2017) also attempted to identify dropout causes but employed a different methodology over more recent scholar census data (2014–2016) from the states of Ceará and Sergipe, Brazil. Two other variables were highlighted as important to dropout detection, namely: teaching-learning model and infrastructure availability (computer labs existence).

Howard and Mozejko (2013) intended to assess the results of the Australian Commonwealth Governments Digital Education Revolution in New South Wales (DER-NSW), specially regarding the one-to-one laptop program. Three online surveys were applied to students of all New South Wales public secondary schools during 2012. The authors were willing to monitor the educational changes caused by the laptops introduction on the school life of students and teachers. The statistical report, among other findings, verified an increase in student-centered practices and students engagement due to the laptop use. Also in Australia, Alom and Courtney (2018) analyzed student's trajectory from primary school until the end of secondary school/university enrollment. The data provided by the Australian Bureau of Statistics revealed several gender disparities – around 10% in some states – considering school year completion.

The factors that impact the academic achievement in junior high school were also investigated recently by Chowa, Masa, Ramos, and Ansong (2015). The work used data from the Ghana YouthSave Experiment, containing information of 4993 students from 89 schools across eight regions of the African country. The data included educational, health, psychosocial, and financial information. The study revealed that class size and presence of toilet facility are helpful predictors for students performance in Mathematics. However, they also identified that student-level characteristics usually play a more important role than school characteristics when it comes to the academic achievement. Pursuing these same goals, Martinez Abad and Chaparro Caso López (2017) applied ENLACE surveys to 18,935 high school students from Baja California, Mexico. The responses were classified using decision trees that allowed the authors to conclude that personal factors are the most relevant ones to the academic performance, followed by school-related and social factors.

In Hutt et al. (2018) used a national dataset of 41,359 college applications for prediction of 4-year graduation rates using sociodemographic and academic data. The authors had the goal to determine the accuracy of predicting college success from application data and investigate if incremental extracurricular and work experiences could interfere on the student performance. The results showed that machine learning algorithms can be used to predict 4-year graduation for over 70% of students using this type of data. At last, Shea and Bidjerano (2014), the authors used a nationally representative sample of data (more than 18,000 students) to determine if students enrolled in distance education courses during their first year in a community college tend to complete a degree with lower rates compared to students who are not enrolled in such courses. Students who take some of these online distance courses have better chances to obtain a community college credential. Table 1 summarizes the main characteristics of the related work presented here.

3. Secondary education in Uruguay

The Uruguayan secondary school lasts 6 years and it is divided into two cycles (see Figure 1). The first one is the Basic Cycle (*Ciclo Básico*), where this work is focused on. It lasts from the first year to the third and is equivalent to Seventh, Eighth and Ninth grades in United States educational system.

In 2006, there was a reformulation on the educational system. The goal was to focus on the learning (*aprendizaje*), instead of instruction (*enseñanza*), thus the student should be evaluated not only by his performance but also by his behavior. In this sense, during the student's evaluation, some items like interest, attitude and social integration are also taken into account. Moreover, the evaluation of the performance is connected with how much the student achieved his proposed objectives (Nahum, 2008).

This reformulation allows a student to surpass to the next year even if he didn't achieve the necessary qualification in up to three subjects. So, he will have a special assistance, during the next year course, containing directed studies which will help him to overcome the past difficulties. The students with insufficient qualification in 50% of the subjects (according to the quantity of

Table 1. Related works and its characteristics.

Work	Scope	Size of data	Source	Goals	Methods	Education level
Bezerra et al. (2016)	Regional	2,699,350 registrations	Brazilian census	Analyze the evasion in public schools	Decision trees/ Association rules/ Logistic regression	Secondary school
Calixto et al. (2017)	Regional	Not informed	Brazilian census	Identify the factors that lead to school evasion	Association rules/Logistic regression	Secondary school
Hutt et al. (2018)	National	41,359 Applications	Common App/National student clearinghouse/National center for educational statistics	Predict 4-year bachelors graduation in a generic manner	Random forest/Hill-climbing	High school/ Graduation
Alom and Courtney (2018)	National	263,413 students	Australian bureau of statistics	Track students from primary school to enrollment in university	Statistics/machine learning methods	Primary and secondary school and graduation
Howard and Mozejko (2013)	Regional	19,748 students/ 2,831 teachers	Australian bureau of statistics	Evaluate the results of the DER-NSW one-to-one laptop program	Statistical analysis	Secondary and central schools
Martinez Abad and Chaparro Caso López (2017)	Regional	18,935 students	ENLACE tests and context surveys	Detect factors linked to academic achievement	Decision trees	High school
Chowa et al. (2015)	Regional	4,993 students	Ghana YouthSave Experiment	Examine the influence of student and school factors on academic achievement	Statistic methods	Junior/High schools
Shea and Bidjerano (2014)	National	16,100 students	BPS 04/09	Examine the associations between enrollment in distance education courses and degree attainment	Binary logistic regression analysis	Post-secondary/ Graduation

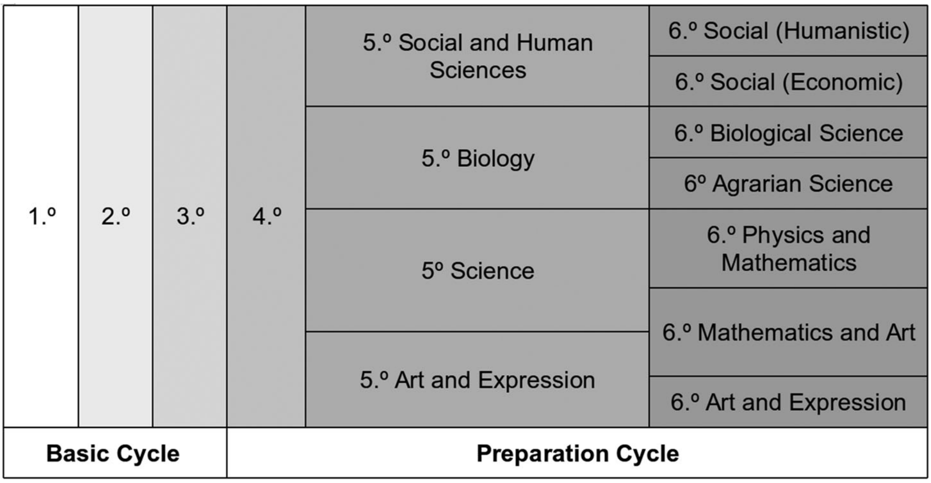
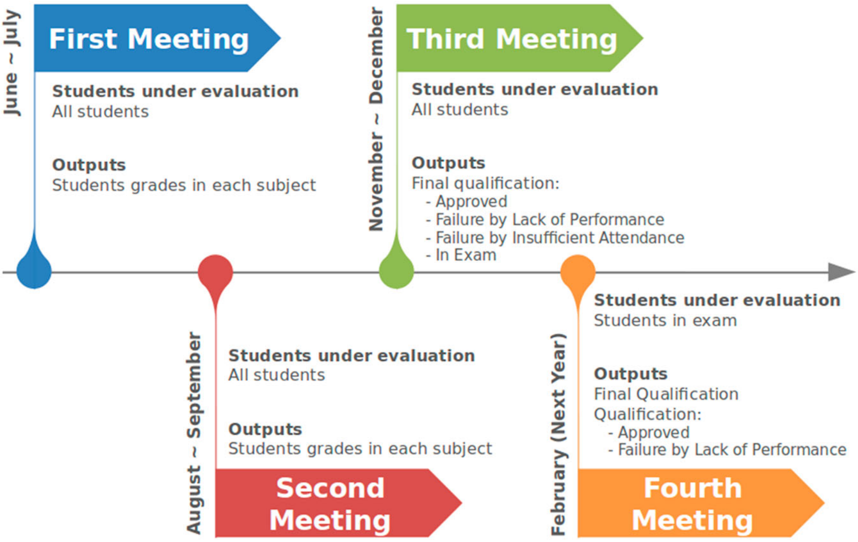


Figure 1. The scenario on the Uruguayan' secondary school and all the available options of courses.

subjects that they are enrolled during the school year and the ones that he failed in the past years) should go to exams. The students that had insufficient qualification on more than 50% of the subjects will fail (Nahum, 2008).

The evaluations take place on the so-called qualification meetings. There are at least three meetings during the school year, when the performances of the students on each subject are assessed and receive a qualification (grade). After the third meeting, the student receives his final qualification



General Rules

Final Qualification	Number of qualifications below grade 6	Number of non-attendances
Approval	Up to 3	Any
Exams	Between 4 and 6	Up to 35
Failure by Lack of Performance	More than 6	Up to 25
Failure by Lack of Attendance	More than 3	More than 35

Figure 2. Overall flow of the evaluation process.

(*fallo* in Spanish), which represents the result of his performance during the school year. The final qualification is categorized as follows: "approved", "failure by lack of performance" (*repite por rendimiento*), "failure by insufficient attendance" (*repite por inasistencia*), or "in exams" (*fallo en suspenso*). In this last case, a fourth qualification meeting will take place after a set of final exams to decide whether the student will finally be approved or not (Nahum, 2008). Figure 2 shows the overall flow of the evaluation process.

Uruguay participated in the last six editions of the Programme for International Student Assessment (PISA), and although the country has shown some improvement in the student's performance in the last PISA (2015), the country still struggles with a performance below the average among OECD countries. Uruguay is ranked at the 48th place among the 72 participant countries, and at second among Latin American countries (Chile is the first). According to the PISA report of 2015 (OECD, 2015), the percentage of Uruguayan students who have repeated a grade during primary and secondary is one of the highest among OECD countries. Moreover, 15-year-olds students in Uruguay performed below OECD countries average in the three main dimensions assessed: science literacy, mathematics, and reading. In science literacy, 15-year-old students in Uruguay scored 435 points compared to an average of 493 points in other OECD countries (boys performed 9 points better than girls). In mathematics, Uruguayan students scored 418 points compared to an average of 490 points in OECD countries (boys performed 14 points better than girls). At last, in reading, Uruguayan students scored 437 points compared to an average of 493 points in OECD countries (in this indicator, girls performed 23 points better than boys).

In the last few years, some problems associated with grade retention and students dropout are concerning the local authorities. For instance, it is possible to observe a diminishing of 10% of the students in the transition from primary to secondary education, and at the age of 13 years old, a 26% of grade retention and a 3% of dropout. When considering the transition from lower secondary education (1st, 2nd and 3rd years) to upper secondary education (4th, 5th and 6th years) it is possible to observe a decrease of 20% in the proportion of students who are enrolled at the expected grade for their age (54% to 34%, respectively). Moreover, there is an increase of 27% of students that do not attend the educational system at all (INEEd, 2017).

Fernández and Pereda (2010) described a phenomenon called "Educational Disaffiliation", which is closely related to the problems enumerated here. The authors stated that both grade retention and dropout are factors causing educational disaffiliation: a series of decision makings that lead a person to modify his position in the social space while remaining in a vulnerable status.

4. Database access and features

4.1. Database description

The database provided by *Administración Nacional de Educación Pública* (ANEP) contains data from the secondary school in Uruguay. This data covers around 135k students and 8k classes (*grupos*). Also, there are some demographic data like address, department (similar to a province), gender and age. All these data refer to 254 centers (schools and academies) spread over 19 Uruguayan provinces during three years (2015, 2016 and 2017). Regarding the educational data, there are information about the final qualifications and the subjects grades of the students, and the total of non-attendances (justified or not) at the moment of each one of the four qualification meetings. The database also contains some information like localization, department, if the center is in a rural or urban zone, and which region of the country is located. Additionally, there are some text fields containing annotations for each student. Teachers make these textual annotations during the qualification meetings.

4.2. Restrictions to access the data

The ANEP did not deliver the data until it agreed on a series of guarantees and reached joint working arrangements. Apart from being a groundbreaking project, it also involves the treatment of underage students' data and the transference of information from students' trajectory all over the country, both matters of concern from the legal perspective. The institution responsible for the project and the team of researchers needed to treat all personal data to which they had access in accordance with Uruguayan Law number 18,331 of August 11 2008 and Decree number 414 of August 31 2009. All the data available could be used only for the realization of this project and any communication or transfer of the data to third parties needed prior written authorization from the *Consejo Directivo Central* (CODICEN). Moreover, the data provided by ANEP must be hosted on servers within the Uruguayan territory, which will only be accessed by the project researchers (including foreign researchers for development purposes). At the end of the project, the institution and the researchers are forced to remove the provided data from all their physical and logical systems. For this reason, it was necessary to install a server under UDELAR domains to access the database and perform the experiments remotely.

5. Methodology

With the increase of available educational data, it is expected that LA will become a powerful tool to inform and support learners, teachers and their institutions in better understanding and predicting personal learning needs and performance (Greller & Drachsler, 2012). However, there are some difficulties in defining a universal methodology for the LA field. With that in mind, the methodology adopted in the present work involved the experimentation of different EDM methods, always aiming to improve the results in each step. More clearly, it was used as a methodology similar to the one proposed by Hübscher, Puntambekar, and Nye (2007) and known as *Interactive Data Mining*. In that, the authors argue that EDM is an iterative process, where in each iteration the researcher "helps" the algorithm to obtain better results and learns from those obtained so far.

This process can be seen as an evolution/adaptation of the core idea of a Knowledge Discovery in Database (KDD) method: first, there must be an identification of the domain and useful prior knowledge, as well as the main goals of the process from the stakeholders point of view; second, target datasets are selected, cleansed and pre-processed; third, it is necessary to match the previously identified goals to some data mining technique; fourth, the generated models are explored in order to find interesting patterns; and finally the encountered patterns are presented in an apprehensible way to the stakeholders (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). As the goals of the present initiative also involved the development of a prototype for the visualization of students trajectory, the results of the mining process also helped to suggest possible visualizations for the dashboard. Figure 3 depicts the overall steps followed throughout this initiative.

Initially, there were several meetings and discussions so the development team could understand the real needs and goals of the Uruguayan authorities towards this initiative. First of all, it was agreed that any tool or generated rule would not be a fully automated actuator, in the sense that any reports or conclusions should be further judged by some specialist. For example, a failure prediction for a particular student would not be delivered to him directly but to his professor or supervisor. Moreover, during the meetings it was raised a matter of concern towards the danger of stigmatization of poor Uruguayan regions, that is, to be cautious if the data statistically shows some critical pattern in these regions.

Once the database (anonymized and decoupled) was made available, the process of extracting the data began. The information was moved from the base located on ANEP into the *Universidad de la República* (UDELAR) server (PostgreSQL database). Several tools were used to handle and analyze the huge amount of data available. However, the most important one was Pentaho Data Integration (PDI), used during the whole process to perform ETL (**E**xtract - **T**ransform - **L**oad) routines, to generate new tables, to clean the records, among other tasks. Pentaho Data Integration enabled fast and straightforward operations over the PostgreSQL database upon which the data was deployed. At each cycle, new proposals of experiments and educational metrics of interest were discussed and

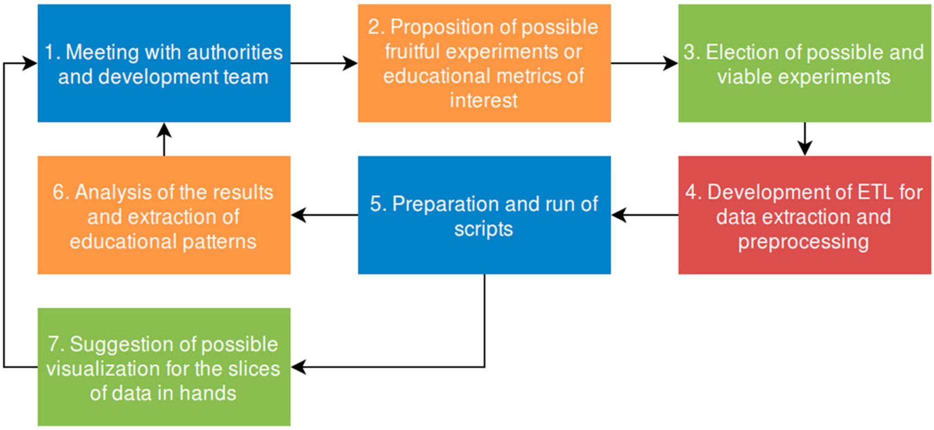

1. Meeting with authorities and development team
2. Proposition of possible fruitful experiments or educational metrics of interest
3. Election of possible and viable experiments
4. Development of ETL for data extraction and preprocessing
5. Preparation and run of scripts
6. Analysis of the results and extraction of educational patterns
7. Suggestion of possible visualization for the slices of data in hands

Figure 3. Adopted methodology.

implemented. The results were then analyzed and presented to the authorities along with suggestions of possible visualizations.

6. Experiments and Results

This section describes the experiments and the obtained results, as well as it indicates all variables in use and how the experiments related to the research goals.

6.1 Clustering students attendance and overall performance

In this experiment, a table containing 17 attributes was generated, where 16 of these attributes were used as input for the model using a clustering algorithm (*k-Means Clustering*), and the variable *fallo* (final qualification) was used to evaluate the clusters (see Table 2). The dataset for this experiment was generated using data from 2015 and 2016, including the student's grades and number of non-attendances in each qualification meeting. The main goal was to find a relationship between the final qualification and the number of subjects where the grades were below a given cutoff on the qualification meetings.

In order to work with performance indicators, we derived variables related to the percentage of subjects in which a given student achieved a certain cutoff grade (from grade 3 to 7). For instance, variables, *psub_ax_my* can be read as follows: *ax* represent the average (cutoff) that is being taken into account with values ranging from 3 to 7, and *my* indicate which meeting the variable refers to with values ranging from 1 to 2. For instance, *psub_a*3*_m*1 refers to the average grade 3 until the time of the meeting 1. That is, it represents the number of subjects (in percentage) in which the student's performance was assessed with a grade lesser than 3 until the moment of meeting 1. Thus, *psub_a*3*_m*2 refers to the same average until meeting 2; *psub_a*4*_m*1 represents the average

Table 2. Variables used and its description.

Variable	Description
qty_subjects	Number of subjects the student attended
age	Student's age
psub_ax_my	Percentage of subjects (according to the total that the student was attending) which the student obtained its qualification (average) less than *x* in meeting *y*, where *x* ranges from 3 to 7, and *y* varies from 1 to 2
injust_na_mx	Quantity of unjustified non-attendances until meeting *x*
just_na_mx	Quantity of justified non-attendances until meeting *x*

grade 4 through meeting 1, and so on. The variables *just_na_mx* and *injust_na_mx* follow the same pattern for justified and unjustified non-attendances, respectively, where *mx* indicates which meeting is being considered and *na* indicates the number of non-attendances.

The implementation of *k*-Means Clustering was done in Java, using Weka API (Hall et al., 2009). We chose to use this tool for its portability, ease of use and extensive documentation. The number of clusters used was $k = 3$, where each cluster represents a possible category of the final situation of the student (approved, failure by lack of performance, and failure by lack of attendance). The initialization of the algorithm was done randomly and the Euclidean Distance was used as a measure of distance.

As it can be seen in Tables 3–5, students from the first and second year of the Basic Cycle are enrolled (variable *qty_subjects*) on average in approximately 12 subjects, and students from the third year in approximately 13. In relation to age, it can be observed that the average age increases according to the final qualification under analysis. Approved students have a lower average age than those who fail by lack of performance, which in turn have lower average age than those who fail by lack of attendance.

According to Table 3, it is verified that until the moment that the second meeting happens, few students have an average grade lower than 5. However, this number increases when average grade 6 is taken into account (6 is the minimum grade required for students approval in secondary schools of Uruguay for the plan of 2006). The number of students with an average of less than 7 at the time of the first meeting increases more sharply, decreasing at the second meeting as they are approved at the end of the school year.

It is important to point out that the number of non-attendances (justified and unjustified) of the approved students is the lowest from the three groups of students (approved, failed by lack of performance, and failed by lack of attendance) – which is obviously expected, but the centroids shed light into these more precise numerical differences.

Looking at Table 4 related to students that failed by lack of performance, it is verified that up to the moment of the second meeting approximately 25% have average grades smaller than 5. However, this value jumps when the minimum average grade 6 is taken into account. At the time of the first and second meeting, approximately half of the students are below the average grade. Still, about 80% of the students have average grades lower than 7 in both meetings.

Regarding the non-attendances, the amount is higher when compared to the values of the approved students. There is a significant increase in unjustified non-attendances recorded at the time of the second meeting.

Table 3. Resulting centroids for approved students.

Variables	First year	Second year	Third year
qty_subjects	11.9583	11.9661	13.0005
age	12.8453	13.8631	14.9177
psub_a3_m1	8.0574	7.9251	5.7562
psub_a3_m2	5.7074	5.6952	4.0129
psub_a4_m1	8.2028	8.1311	5.984
psub_a4_m2	5.8289	5.8404	4.1947
psub_a5_m1	8.9928	9.0428	7.0879
psub_a5_m2	6.448	6.5154	5.0949
psub_a6_m1	14.1091	14.8976	13.5926
psub_a6_m2	10.1624	10.6905	9.8643
psub_a7_m1	35.3271	37.2662	37.4944
psub_a7_m2	24.0937	25.7453	26.4785
injust_na_m1	2.4528	3.4541	3.9361
injust_na_m2	5.809	7.6511	8.9487
just_na_m1	1.0515	1.1833	1.3963
just_na_m2	2.6067	2.9512	3.5368

Table 4. Resulting centroids for students that fail by lack of performance.

Variables	First year	Second year	Third year
qty_subjects	11.9687	11.9698	12.994
age	13.3956	14.339	15.3671
psub_a3_m1	11.4198	11.2395	9.0989
psub_a3_m2	9.3592	9.0174	7.7194
psub_a4_m1	15.1444	14.7945	12.9739
psub_a4_m2	14.4877	13.4724	12.5046
psub_a5_m1	26.1789	25.4489	24.1916
psub_a5_m2	26.3684	24.583	24.1661
psub_a6_m1	52.3576	50.9826	50.3718
psub_a6_m2	50.9396	48.7611	48.6528
psub_a7_m1	83.0485	81.656	81.6131
psub_a7_m2	79.1623	77.4063	77.5701
injust_na_m1	5.8863	6.9282	7.1282
injust_na_m2	13.5609	14.7964	15.6168
just_na_m1	1.764	1.9489	2.3047
just_na_m2	4.4637	4.8509	5.3797

Table 5 presents the data of the students who failed by lack of attendance. It is possible to notice that already in the first meeting, approximately half of the students have an average grade lesser than 4. The same amount of students gives a jump in the second meeting. If the average grade of 6 is taken into account (minimum grade for approval), more than 80% of students are below this already at the time of the first meeting. In relation to the average grade 7, it is verified that more than 95% of the students are below this average already in the first meeting, in all three years.

Also, as these students failed by lack of attendance, it can be seen that the number of unjustified non-attendances in first and second meetings (*injust_na_m*1 and *injust_na_m*2) are quite high, especially when compared to the values of the other two previous groups. It is also possible to verify that the justified non-attendances (*just_na_m*1 and *just_na_m*2) do not present great differences in relation to the other groups. From these data, it is possible to identify, with some confidence, the students who are at-risk of failing by lack of attendance in advance.

One can also compare the average profile of the student of each category in a given year (inter-table comparison). For example, it turns out that there is no significant difference between the average profile of the approved student and the student who fails at the bottom. For this, only the low cut-off variables (up to 4 – *pmat_m*3 and *pmat_m*4) should be taken into account in any of the years (comparison between analog columns of Tables 3 and 4).

Table 5. Resulting centroids for students that failed by lack of attendance.

Variables	First year	Second year	Third year
qty_subjects	11.9635	11.9689	12.9838
age	14.5923	15.231	16.1978
psub_a3_m1	43.0403	37.6328	36.0106
psub_a3_m2	67.6904	60.3698	58.3756
psub_a4_m1	56.0875	50.1981	48.4257
psub_a4_m2	79.1243	72.5361	70.5735
psub_a5_m1	72.5946	67.207	65.6736
psub_a5_m2	88.7003	84.0458	82.7778
psub_a6_m1	88.5776	84.9484	84.1
psub_a6_m2	95.6351	93.2301	92.4206
psub_a7_m1	96.9901	95.6756	95.1195
psub_a7_m2	98.7948	97.8617	97.6163
injust_na_m1	20.2722	18.9524	17.9557
injust_na_m2	52.2824	48.2996	47.0096
just_na_m1	2.67	3.2957	3.7215
just_na_m2	4.4413	5.8874	6.3571

However, this difference begins to become noticeable from the *pmat_m*5 variable. That is, for the cut-off point in average grade 5: while the students approved in the first year have about 9% of the subjects below 5 in the first meeting, the student who fails by lack of performance accounts for 26% of the subjects below of average grade 5. It is also possible to realize that for any of the years, the unjustified non-attendances variables are the most unequal when comparing the three groups of students.

Some of the patterns found using k-Means Clustering for identification of students at academic risk are:

- Students with five unjustified non-attendances;
- 50% of the student's grades below five at the moment of the first evaluation meeting;
- Five or more unjustified non-attendances in the moment of the first evaluation meeting, and twice this value in the moment of the second evaluation meeting;
- A student with the age above the average age on a given year (that had failed at least once), and with 50% of its grades below six in the moment of the first evaluation meeting.

Even though the use of k-Means clustering algorithm achieved low accuracy for this experiment (mostly if compared with classical classification algorithms such as Neural Networks), it provided readable results that can be used as possible educational metrics. The technique fit the initial purposes and served as an alternative approach to perform an exploratory descriptive analysis of the data. Figure 4 shows the clustering results (centroids denoted by X and real students by circles) presented to the user after the selection of a target student (red triangle). Each cluster is assigned to a percentage representing an academic risk – computed by the number of students who were not promoted in that cluster divided by the number of students in that cluster. This visualization will be further integrated in the dashboard under development (see Subsection 7.1).

6.2 Clustering the explicit performance of students

In this second experiment, we used only the subjects grades along with the attendance data (raw data available in the database). This way the clusters could help on mapping specific subjects (such as Math, Spanish, etc.) whose performances have a deeper impact on the students final qualification.

When the number of clusters was equal to two (i.e. treating as a binary classification – failure or success) the results simply reflected the educational rules of *Consejo de Educación Secundaria*

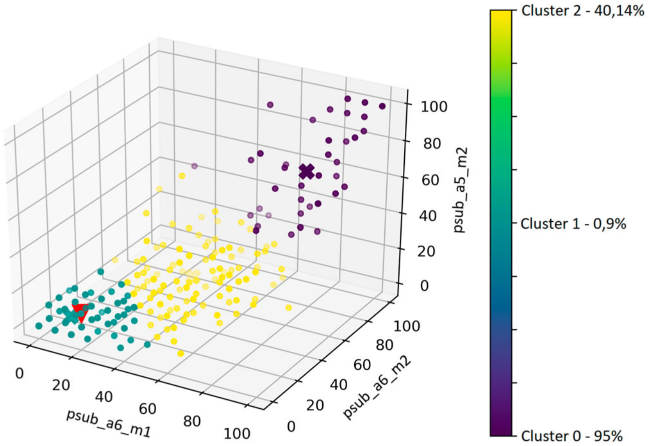


Figure 4. An example of visualization to evaluate the academic risk.

(CES) – upon which a student will fail on a subject if achieves a grade lesser than six at the third evaluation meeting and will fail on a course if fails on more than six subjects. The experiments also generated other rules that are not detailed here for simplicity and space constraints (for instance, related to retaking exams). Adding up more clusters, interesting patterns emerged. For instance, when $k = 4$, considering the first year of Basic Cycle, the algorithm detected intermediate groups of students. That is, although there were two well defined groups – C1 and C2 – with dramatic (for better or worse) grades and non-attendances values, there were also other two groups – C0 and C3 – of students who appear to be on the borderline of the failure/success outcome. Their differences with respect to the subject grades are almost indistinguishable; however, the impact of different non-attendances rates can be clearly seen among the distinct groups as shown in Table 6 (some subjects were omitted to ease readability).

The main findings of this experiment are:

- These initial experiments suggested a very homogeneous distribution among the grades of all subjects inside each cluster. These grades and their respective final outcomes just reflected the rules detailed at 2;
- However, as the number of clusters was increased, clusters of students who stand at the borderline of failure/success were revealed;
- Although further experiments should be conducted with greater number of clusters, these two borderline clusters (C0 and C3) differ specially regarding their average of non-attendances – the cluster with positive outcome has half the non-attendances of his opposite, at both meetings.

These four clusters were built upon data from 2015 to 2016. We then used data from 2017 to evaluate the clusters by placing new students on this *n*-dimensional clustered space and verifying which centroid is the closest one to a given student. The results were very accurate: 99% of the 2017 students were correctly classified. However, as mentioned before, the sections made by the clustering algorithm, specially on the subjects grades, are quite well-behaved among different subjects and no conclusions can be drawn about how it impacts on the final outcome.

6.3 Using association rules to find troublesome subjects

A different approach was used to analyze subjects that interfere in the school performance of a student enrolled in the educational system of Uruguay. The *Apriori algorithm* was used to perform a two-step analysis: first, a set of frequent items is created; then, association rules are inferred from that set. The approach used was efficient, since besides it produced good results, it can be easily understood, even for people with no experience in data mining.

From the analysis of the most frequent items in dataset, the first evaluation to be made refers to the subjects that most lead students to exams. Figure 5 presents these results, wherein the *x*-axis support is shown for each subject. This measure represents how often an item appears in the dataset (Tufféry, 2011). The higher, the more frequent the item.

Table 6. Detailed results (as centroids values) of 4-means clustering algorithm.

Variable[a]	C0	C1	C2	C3
Non-attendances	8/19	21/55	2/5	4/9
Spanish Lang.	5/5	3/2	8/8	6/6
History	5/5	3/2	8/9	6/6
Mathematics	5/4	3/2	8/8	6/6
Biology	5/5	3/2	8/8	6/6
Outcome	Fail	Fail	Success	Success

[a] Values separated by slashes mean: value at the first evaluation meeting/value at the second evaluation meeting.

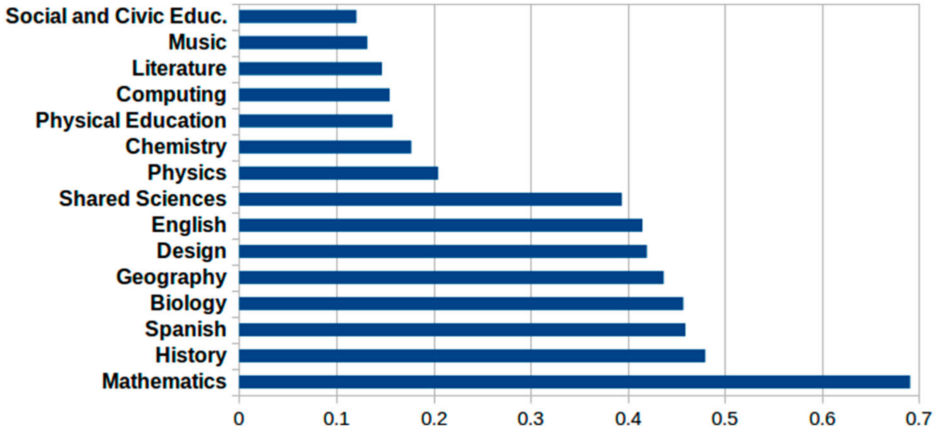


Figure 5. Subjects that most lead students to an exam situation.

It can be observed that students have the greatest difficulty with Mathematics, having a relatively large difference in relation to the second most frequent subject (History). This, together with Spanish, Biology, Geography, Design, English and Shared Sciences[1] have a relatively similar frequency. The last seven subjects also have a similar frequency.

Figure 6 shows which are the fifteen combinations of subjects that appear most frequently in the data set, where the axis *x* presents the support of each combination. It can be seen that Mathematics appears in the seven most frequent combinations of dataset, endorsing the results presented in Figure 5. History, which appears in the most frequent combination associated with Mathematics, is present in four more combinations; as well as Biology, appearing five times in these results. It is also noticed that Spanish appears frequently, evidencing issues regarding the country native language. It can be verified that the majority of the students that go to exams are later approved (after the provision of the exams). Still, most students who go to exams have more than 25 unjustified non-attendances. In addition, this analysis showed that in 2015 and 2016, most students were enrolled in the third year, followed by second and first years, with fewer students.

In the generation of association rules (second analysis), the student's final qualification (*fallo*) was used as a consequence and 91 rules were obtained. Table 7 displays these results, ordered by *lift*.

Table 7 displays the first fifteen association rules generated, ordered by the *lift* value. It can be seen that in these fifteen items the consequent is a "fail by lack of performance". Using the first rule as an

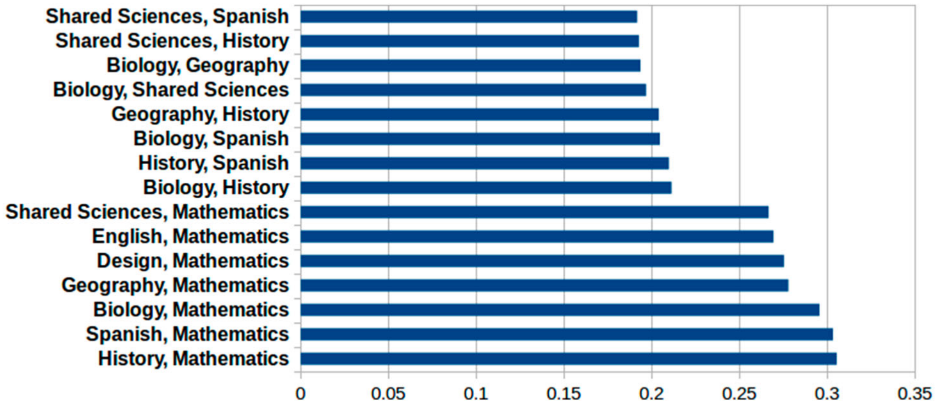


Figure 6. Combination of subjects that most lead students to an exam situation.

Table 7. Rules of association generated using the Apriori algorithm.

Antecedent	Consequent	Confidence	Lift
Spanish, Mathematics, English	Failed by lack of performance	0.439	1.338
Shared Sciences, Biology, Mathematics	Failed by lack of performance	0.432	1.316
History, Mathematics, Biology	Failed by lack of performance	0.429	1.306
Spanish, Mathematics, Biology	Failed by lack of performance	0.422	1.288
Mathematics, More than 25 unjustified absences	Failed by lack of performance	0.422	1.285
Mathematics, English	Failed by lack of performance	0.420	1.280
History, Shared Sciences, Mathematics	Failed by lack of performance	0.420	1.280
History, English	Failed by lack of performance	0.419	1.277
Shared Sciences, English	Failed by lack of performance	0.418	1.274
Biology, English	Failed by lack of performance	0.418	1.274
Spanish, Shared Sciences, Mathematics	Failed by lack of performance	0.417	1.271
Spanish, Mathematics, History	Failed by lack of performance	0.417	1.270
Geography, English	Failed by lack of performance	0.414	1.263
Spanish, English	Failed by lack of performance	0.401	1.221
Design, English	Failed by lack of performance	0.396	1.207

example, one can interpret that the learner in Spanish, Mathematics and English ends up being reproved by lack of performance after the exams are given.

Again, Mathematics is quite common in the rules, appearing in the first seven items, and in nine of the fifteen presented. Besides Mathematics, English appeared as a very frequent subject, being present in eight of the fifteen rules generated. An interesting fact is presented in the fifth rule, where it is shown that in student who goes to exams in Mathematics have more than 25 unjustified absences. This may point out that students who do not attend classes tend to have more difficulties, in this case, associated with Mathematics. Alternatively, this may happen because the student has difficulties, and is not motivated to go to the classes.

Table 8 presents the generated rules for the students who have the final qualification "approved", ordered by lift. It can be seen that this measure is smaller than those presented in Table 7. This is because the number of students who have been approved is higher than those who have been failed. The main findings of this experiment are:

(1) Mathematics is the subject that most lead students to exams, appearing on almost 70% of the exams situation, and also associated with other subjects (top seven associations);
(2) Spanish, the native language, is also a source of problems for Uruguayan students; when combined with Mathematics, appears on more than 30% of the exams situations;
(3) History is another subject that brings problems to the students: "History and Mathematics" is the combination that most leads students to exams (more than 30% of the cases);
(4) Biology is also a troublesome subject: combined with Mathematics leads almost 30% of the students to exams;
(5) The number of non-attendances can bring some insights about the student's situation since when approved, students tend to have up to 15 unjustified non-attendances;
(6) The students who fail by lack of performance tend to have more than 25 unjustified non-attendances.

Table 8. Association rules generated for students that were approved.

Antecedent	Consequent	Confidence	Lift
Mathematics, between 6 and 10 unjustified non-attendances	Approved	0.711	1.058
Design, between 11 and 15 unjustified non-attendances	Approved	0.707	1.052
Geography, between 11 and 15 unjustified non-attendances	Approved	0.699	1.040
Biology, between 11 and 15 unjustified non-attendances	Approved	0.688	1.025
Design, between 11 and 15 unjustified non-attendances	Approved	0.684	1.018
Spanish, between 11 and 15 unjustified non-attendances	Approved	0.677	1.008
History, between 11 and 15 unjustified non-attendances	Approved	0.677	1.007

7. Discussion

The development of a countrywide Learning Analytics initiative involves a series of challenges and difficulties that need consideration and are worthy to mention. For instance, many works in the literature of Learning Analytics and Educational Data Mining revolve around data collected by the very researchers who are conducting the reported experiments. That being said, it is expected that in those cases, the team responsible for the technical development of the project and the scientific experimentation are quite aware of the specifics related to the data they are using, the educational contexts and problems they are dealing with, as well as the important variables at stake. It is also expected some sort of control about the directions the project will follow. This situation is entirely different when one is dealing with countrywide data that was made available for the first time, for the sole purpose of the ongoing initiative, and with a number of legal constraints requiring attention. This section intends to answer the research questions elaborated in Section 1 at the same time it discusses the main findings of the initiative so far and some of its limitations.

RQ1 – Which are the main challenges in order to implement a countrywide Learning Analytics initiative?

Table 9 summarizes some of the most important challenges faced in this initiative, together with suggestions for solutions. These challenges were faced during the implementation of the project as a whole and can be classified into three main categories: bureaucratic, technical and methodological.

Dealing with national data requires an in-depth investigation of the country educational system, and its difficulties and demands (Challenge 1 in Table 9). Moreover, the identification of the most important factors affecting the main educational problems of an entire nation is not an easy task and it requires the involvement of a multidisciplinary team, but also the support and close engagement of governmental actors. This was, fortunately, the case of the present initiative, that was carried out by a group formed by professionals with different backgrounds (computer science, communication, philosophy, law and education) and that counted with the strong support (both technical and pedagogical) of a governmental actor with deep expertise on the educational system of the country.

In order to successfully conduct such an initiative, it was necessary to cope the initial proposal approved by the funding agency (and the initial expectations of the researchers) with the needs and inputs of the governmental actors engaged in the process (Challenge 6 in Table 9). For instance, the initial idea focused on the development of models to detect at-risk students was redefined, giving place to the development of a prototype focused on tracking students trajectory together with the provision of general indicators for decision making. The initiative shifted the initial view of academic risk to the detection of success factors and good practices, at the same time that started to consider with much more attention the avoidance of stigmatization and possible bias towards negative factors. Such changes required a methodological redefinition that demanded time and directly interfered in the agenda. Shifting the goals from prediction to the provision of general indicators for decision making required to choose mining techniques able to generate human-comprehensible results. For instance, even though a given technique for classification presents a well-known achievable performance, the researchers needed to prioritize the clearness of the discovered patterns in order to provide useful indicators.

Working with sensitive data was another challenge faced during the execution of the initiative. Official negotiations to access the database from ANEP started immediately after the project approval and took much more time than initially planned. A number of technical, ethical and legal requirements needed to be satisfied beforehand (Challenge 2 in Table 9). Security measures were adopted to warranty the confidentiality of personal data, together with measures to avoid data adulteration, loss, consultation or unauthorized access, and to detect data leakage. Experiments were all carried out remotely on a server located at Uruguay, as the international partners were not authorized

Table 9. Challenges and solutions suggestions.

Challenges	Category	Suggestions
C1: Understand the different nuances of the educational system	Methodological/ Bureaucratic	1. To perform a deep study about the structure and functioning of the educational system; 2. to incorporate educational experts and people responsible for the design and implementation of the informational systems currently in use
C2: Satisfy all ethical and legal requirements without narrowing down the scope and potential novelty of the project	Bureaucratic/ Methodological	1. To consult legal authorities in advance to better understand possible ethical and legal requirements and project constraints; 2. to assure security and confidentiality of the personal data throughout the entire project; 3. to implement measures to avoid data adulteration and loss, as well access from unauthorized personal
C3: Time delay to get the signatures and the approval to access the data	Bureaucratic	1. To allocate time in the project schedule for such an activity; 2. to consider in advance the pitfalls of possible delays and to include alternative activities that can be carried out in parallel; 3. to find a balance between the time necessary for this activity and the time needed to conclude other parts of the project.
C4: Dealing with integration and inconsistencies in the databases	Technical	1. To involve technical experts able to identify and rectify any possible data inconsistency, as well as to understand the whole datasets "picture" in order to properly integrate them; 2. to check the existence and completeness of data and systems documentation during project preparation
C5: The systems were not developed to the end of the project	Technical	1. It is necessary to properly use several data science tools and software to organize and process all data; 2. again, it is necessary to count on professionals/ researchers who can deal with this kind of problem: there should be know-how on database integration and anonymization, but also on the preparation of different kinds of data to feed the available algorithms or tools; 3. to consider further training with the users of the systems focusing on the information they must provide in order to allow the LA system to work on the future
C6: To cope the initial proposal approved with the needs of the agency	Methodological/ Bureaucratic	1. To schedule project meetings in which all the stakeholders are able to state their views and needs regarding the project idea. In our project, these meetings helped on the decision of changing the initial goals towards the development of a prototype focused on tracking students' trajectory and the provision of general indicators for decision making, instead of focusing only on academic risk
C7: Identify and select the best alternatives regarding algorithms and machine learning tools	Technical/ Methodological	1. To consider other aspects during the model (algorithm) selection other than performance, such as clearness and readability
C8: Transfer the obtained results to the visualization tool	Technical	1. To show machine learning algorithm results in an apprehensible way to educational professionals

to transport or transfer the data to locally work with it. All these constraints reduced the time available to dedicate to the developments of the experiments, and lead the researchers to make a special ask for the funding agency to extend the project deadline in 6 months. The initiative initially intended to work with data from primary and secondary education of Uruguay. The delay to access the data (Challenge 3 in Table 9), together with the need of integrating databases (Challenge 4 in Table 9) developed under different backbones (with no *a priori* matches between records of each database) imposed another restriction to the initiative that had to restrain its scope to the Secondary Education level only. At last, is it highly important to consider that the systems in use and that collect students data was not developed to the end of the goals presented here (Challenge 5 in Table 9). For instance, the current systems that collect students' presences (specifically regarding the CES database) were not developed to be used in a scenario as proposed by the project and the attendance data may not be fed in a constant manner. This may cause problems when providing data to any tool that demands some level of data reliability and cohesion. More than that, it may imply in some re-education of the users of these systems in order to foster a different behavior while feeding the systems with information.

RQ2 – How feasible is to uncover useful educational metrics inside a countrywide database?

The present project was able to unveil educational metrics that can be used by the authorities for planning educational policies and actions. However, it is important to mention the path to achieve such results was not straightforward and involved a number of methodological decisions (some of those were already mentioned before) (Challenge 7 in Table 9). For instance, Secondary Schools in Uruguay apply more than 15 different curriculum plans that are heterogeneous with respect to several aspects (rules for students' approval, required subjects, number of qualification meetings, etc.). Therefore, the inclusion of all these plans in the experiments would have caused an exponential increase of complexity to data mining applications as well as to the visualization tools (Challenge 8 in Table 9). Considering that, and in order to warrant the feasibility of this first initiative, it was necessary to search for a homogeneous and reasonably encompassing scope. So, we elected for the experiments and for the prototype only the most representative plan in the database (the curriculum plan that covered 82% of the data) in detriment of the less representative ones.

In this work, two distinct types of results were unveiled. The first one has to do with how performances were distributed among students who were approved or failed and considering two aspects: (1) how many qualifications (percentages) were below a certain threshold; and (2) considering students' qualifications themselves in specific subjects. The second result is related to the most troublesome subjects (the ones that most led to exams) to certain students. The feedback given by ANEP authorities on these two results was positive. While the prior results statistically confirmed some suspicious of the authorities and are useful to detect at-risk students and act on early interventions, the former can serve to the establishment of countrywide policies focused on specific topics that need attention (for instance, building summer programs focused on the most problematic subjects). These metrics, although potentially useful, are nevertheless limited by the small amount of qualitative information present in the CES database – there are only non-attendance and subjects grade data in a significant amount. The data about students specific aptitudes, difficulties and disabilities are so sparsely populated we chose to not include in our analysis.

7.1. Learning analytics dashboard prototype

Learning analytics dashboards have been mentioned as highly important to provide visualizations of several aspects related to the teaching and learning processes (Duval, 2011; Verbert, Duval, Klerkx, Govaerts, & Santos, 2013). The main idea is to present big volumes of educational data in a way they can make better sense to the different stakeholders involved (teachers, students, administrators, researchers) and according to their needs. For instance, they can help professors to follow learning process, identify at-risk students, or to better understand how different educational resources are being used in their disciplines (Einhardt, Tavares, & Cechinel, 2016).

Learning Analytics dashboards have been developed with different functionalities, using a great variety of data sources and targeting distinct educational contexts. A research review conducted by Schwendimann et al. (2017) has shown that only 9% of the papers encountered in the relevant literature with respect to learning analytics dashboards cover institutional databases (most of the papers are focused on data logs), at the same time, no mention to countrywide data was made. The authors also highlight that more than half of the papers are focused on university settings and point out to the need of covering other settings such as K-12.

Considering that, the present initiative also fulfills an existing gap in the current literature by delivering a learning analytics dashboard prototype using countrywide K-12 governmental data. The dashboard will be put in use for future validation by the authorities. The different visual modules that compose the dashboard were developed with front-end technologies based on HTML, CSS, JavaScript and using D3.js[2] together with JQuery[3] and Bootstrap.[4] By using the dashboard, it is expected that education professionals and authorities will be able to follow students' trajectories and to contrast educational information across the country. An initial view of the most important visualizations available so far is presented in Figure 7.

The prototype allows to check information according to different categories of graphs. For example, the education professionals can check the grades variation in each year (2015, 2016 or 2017) and school year (first, second, third), besides information about students' approval or failure according to each State, qualifications for each school, among others. Figure 8 shows an example of a more detailed visualization. Here, the user is able to select each region of the country (right graph in the figure) in order to visualize information related to the percentages of subjects below a given grade threshold (up-left graph in the figure), and the number of non-attendances for the different students qualifications (approved, failure by lack of performance or fail by insufficient frequency) in each evaluation meeting of a given year (bottom-left graph in the figure).

Moreover, Figure 9 shows a graph that compares the average grades of students from two schools according to different subjects. It can be seen which one yielded the best results according to each subject. The data can be filtered by year, and schools of a given country region. This kind of visualization helps to evaluate quantitative differences related to performances between schools and to eventually detect strengths and weaknesses of the students in certain domains. For instance, as can be seen in the figure, one of the schools (in blue) outperforms the other in Math, Chemistry, Physics, while the other (in orange) performs better in Music, English, Geography, Biology and Visual Communication.

Figure 10 shows the flow of students by each year according to the final result, reflecting the group trajectory according to the filters used. The data can be filtered by year (first, second or third), country region and school. With this type of visualization, it is possible to have an idea of how the students' performance changes according they advance in the school system. In the example, it can be seen that most of the students that were approved in 2016 had the same result in 2017.

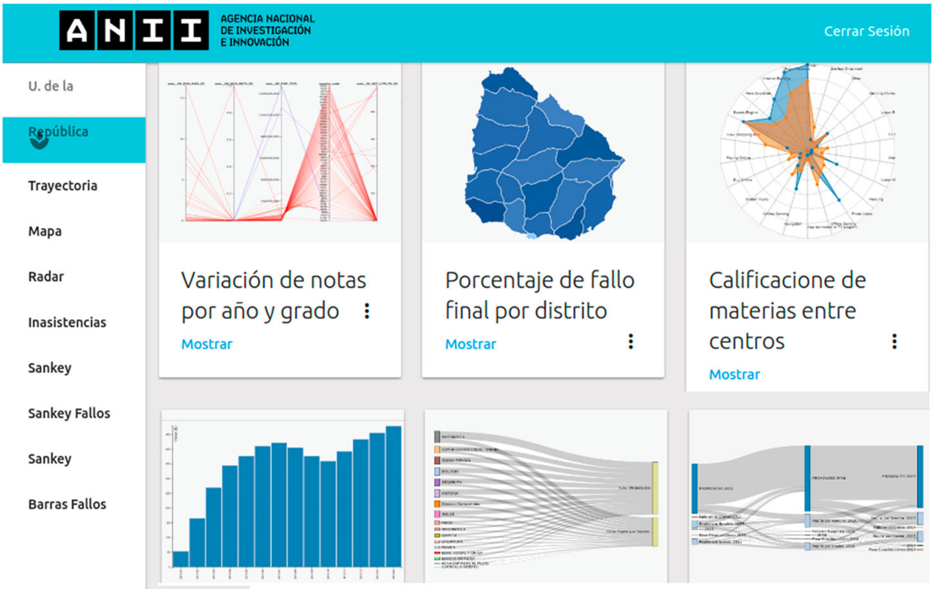
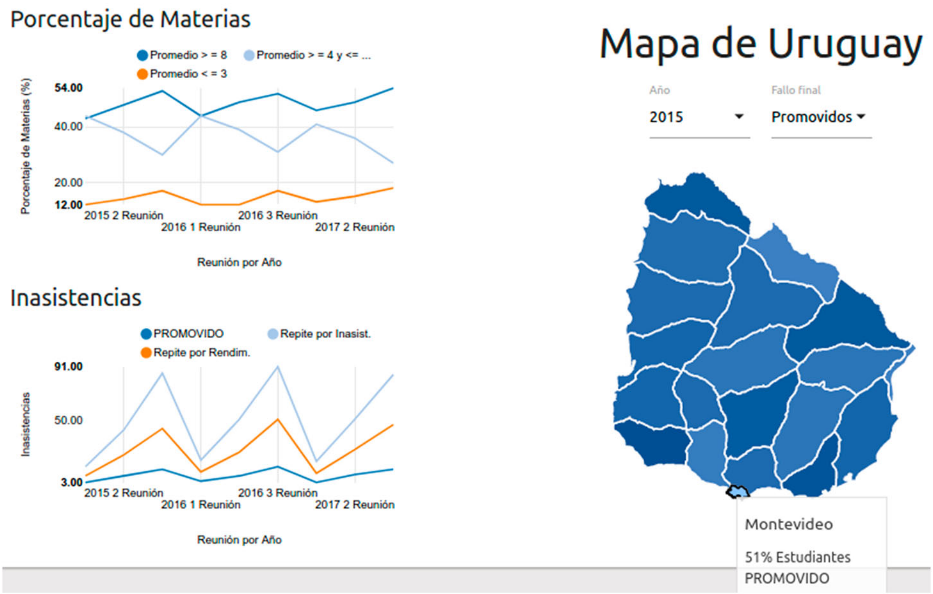


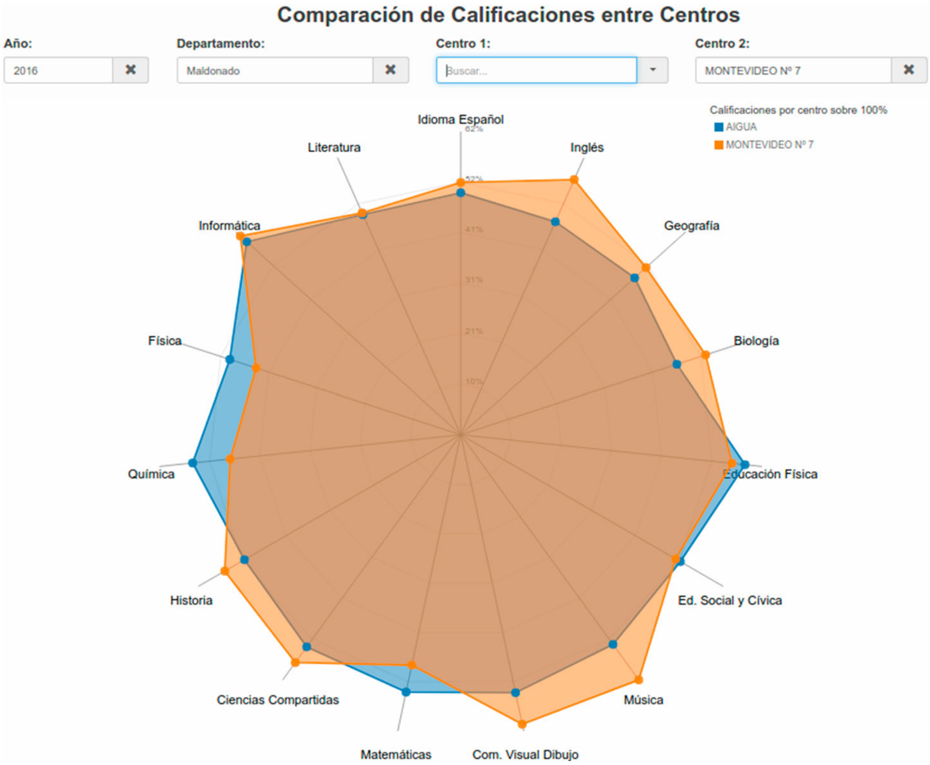
Figure 7. Visualization prototype main screen.



Año 2015

Fallo final Promovidos

Figure 8. Uruguay map - performances and non-attendances in each qualification meeting.



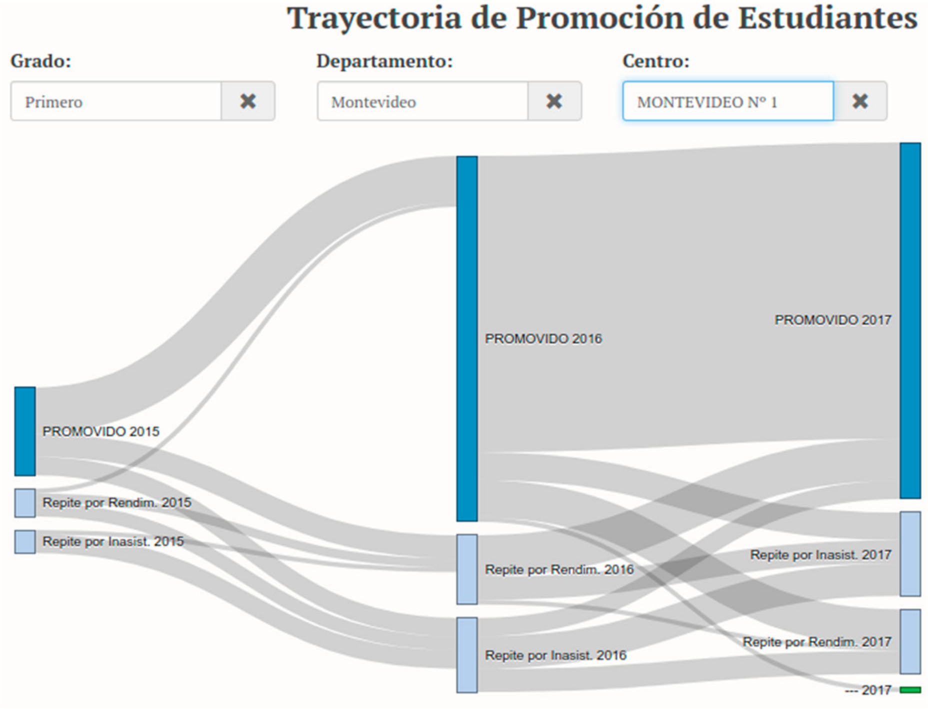
Año: 2016

Departamento: Maldonado

Centro 1: Buscar...

Centro 2: MONTEVIDEO Nº 7

Figure 9. Comparison of performances on different subjects between schools.


Trayectoria de Promoción de Estudiantes

Grado: Primero ✖

Departamento: Montevideo ✖

Centro: MONTEVIDEO Nº 1 ✖

Figure 10. Trajectories according to the final qualification of the students through the years.

8. Final remarks and future work

Developing Learning Analytics studies and tools at national scale involves different kinds of technical, methodological and ethical barriers when compared with smaller and local contexts. This article presented the complexity of this endeavor and the design decisions taken to solve, or work around, the most important issues. The lessons learned during this project are a contribution to facilitate the implementation of other projects at a similar scale. Based on the data obtained from the academic years of 2015, 2016, and 2017, the relationship among different student variables, such as attendance or grades, and performance variables, such as promotion, was investigated. Out of this study, several confirmatory and non-expected rules were found. This analysis helped the academic authorities to better understand the historical data and its value for decision making. It also contributes to insights about the functioning of the Uruguayan early-secondary education.

The most immediate result of this work is a visualization tool that is able to provide support to education professionals to improve their decision making for more than 80% of the Uruguayan public secondary school. This tool visually presents the trajectory of a student through an interactive analytic dashboard. Apart from raw data visualization, the rules found during the clustering experiment (Subsection 6.1) are being used as a basis for more advanced functionality, such as easily finding at-risk students. Although the tools are very tailored to the specificity of the Uruguayan K-12 system, this technical development contributes visualization techniques adapted to longitudinal academic data of any K-12 context.

This article presented the results of a first attempt to make use of a large amount of K-12 information usually store in governing educational bodies. As such, there are several improvements that were proposed after the initial use of the system by its users. The most important are the following:

- After seeing the results of the project, ANEP also made available the primary school database from *Consejo de Educación Inicial y Primaria* (CEIP). However, this database was not explored in this work due to time constraints. It is worth remarking that CEIP database has some particularly interesting information, such as students socioeconomic status and schools infrastructure (not available on the CES database);
- There is also an intention on performing experiments to find correlations between a CEIP trajectory and some outcome metric of the students first year on CES. We also believe that an association rules approach, similar to the one described in Subsection 6.3, could lead to significant results at the identification of how primary school students features (such as writing, reading and communication skills – which are all available in this database) affects their first year on the secondary school;
- In the future, the clusters centroids will also be presented in a 3-dimensional scatter plot together with some selected student (see Figure 4), so the user can visualize where the target student is placed and how close he is to any of these clusters. These clusters will be attached to some level of academic risk; that is, the number of students of a given cluster who failed divided by the total number of students of the cluster. Also, the user will be able to select not only the variables but also the neighborhood of the target student, for instance, to cluster only past students who were at the same school of the target or the same department;
- In the next iteration of the analysis, new derived variables will be included and their effects in the attempt to identify patterns will be evaluated. Another possibility of future experimentation will be based on the distinction of subjects that took the students to an exams situation (and their respective final qualifications) in order to help in the creation of specialized support strategies according to the results. As mentioned earlier, future experiments are also expected to help building a broader notion of student trajectory through the inclusion of data from primary education, that is, to identify relationships between the average profile of a given student and their learning characteristics in secondary education. With the intersection between the two databases, it is expected to contribute to a greater understanding of the evolution of the Uruguayan student within the public education system, in the transition to the first years of secondary education. Also, the visualization prototype will undergo through a validation process by the authorities.

Notes

1. The Shared Sciences subject deals with varied topics. That is, it can address a specific topic in the first year, and vary in the second and third years. However, the statistics of all degrees were saved with this name.
2. https://d3js.org/
3. https://jquery.com/
4. https://getbootstrap.com

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Agencia Nacional de Investigación e Innovación [grant number FSED_2_2016_1_130897].

Notes on contributors

Luiz Antonio Macarini received his bachelor's degree in Computer Engineering (2018) on Federal University of Santa Catarina. Nowadays, he is enrolled on the Post-Graduation Program in Automation and System Engineering (PPGEAS) at the Automation and Systems Department (DAS), Federal University of Santa Catarina (Florianópolis), Brazil. His research is focused on Computer Vision, Machine Learning and Deep Learning. Also, working with Educational Data Mining and Learning Analytics, having some works published in the area.

Henrique Lemos dos Santos is a graduate student at the Institute of Informatics – Federal University of Rio Grande do Sul. He holds a Bachelor's degree (2017) on Computer Engineering from the Federal University of Pelotas. He was one of the researchers involved on the RIURE project (Red Iberoamericana para la Usabilidade de Repositorios Educativos) financed by CYTED (2013–2016), participating in short-time internships at the Escuela Superior Politécnica del Litoral – Ecuador and at the Universidad Nacional de Rosario – Argentina. During his graduation, he was a scholarship student from 2014 to 2017, working on projects about learning objects recommendation and performance metrics of distance learning students at the university.

Cristian Cechinel received the bachelor's and master's degrees in computer science from the Federal University of Santa Catarina, Brazil, in 1998 and 2000, respectively, and the Ph.D. degree in information and knowledge engineering from the Computer Science Department, University of Alcalá, Spain, in 2012. He is currently an Associate Professor with the Sciences, Technologies and Health Center, Federal University of Santa Catarina. His research mostly focuses on the development and analysis of digital learning technologies, learning analytics, and distance learning. He is an Active Member of the Latin-American Community on Learning Technologies and a Former Member of the Special Committee on Computers and Education, Brazilian Computer Society. He is also an Associate Editor of the Brazilian Journal of Computers in Education (2016–2018).

Xavier Ochoa is an assistant professor of Learning Analytics in the Department of Administration, Leadership, and Technology of the Steinhardt School of Culture, Education, and Human Development. He is also a member of the Learning Analytics Research Network (LEARN) at NYU. Xavier holds a Ph.D. in Computer Science from the University of Leuven (Belgium), an M.Sc. in Applied Computer Science Sciences from the Vrije Universiteit Brussels (Belgium) and a B.S. in Computer Science from Escuela Superior Politécnica del Litoral (Ecuador).

Virgínia Rodés coordinates the Virtual Learning Environments Program (ProEVA) and the Open Educational Resources Center (Núcleo REAA) of the Universidad de la República, Uruguay. She has developed research in conjunction with academic networks in Latin America and Europe. In the framework of these collaborations, she has led more than 20 R&D projects with funding from the EU ALFA Program, CYTED, AECID, PASEM, CSIC, FRIDA and ANII, contributing to the development and dissemination of important open and accessible initiatives (LATIn, ESVIAL, BIDYA), in the Latin American region, with her own contributions in the area of technology integration in higher education, access to university study materials, strategies for the adoption of open educational resources, and methodologies for the collaborative creation of open textbooks.

Alén Pérez Casas is an assistant professor at the Department of Academic Technical Support, Sectoral Commission for Education (CSE) and assistant professor of Social Research Methodologies at the Faculty of Information and Communication (FIC) of the University of the Republic (UDELAR), Uruguay. He is a Sociology Magister specialized in qualitative and quantitative research methodologies applied to the study of the social impact of information and communication technologies and Computer Analyst specialized in research and development of information systems applied to higher education.

Pedro Pablo Lucas is a research assistant at the Information Technology Center (CTI), ESPOL University in Ecuador. Pedro received his bachelor's degree in Computer Science from ESPOL in 2015. He has worked on several projects related with real-time systems and artificial intelligence that involved PS4 and PC videogames (2012–2018), algorithmic music composition (2014–2015), swarm intelligence (2016–2017), and augmented reality (2016–2017). Currently, he is working on the development of visualization tools for learning analytics projects and his main focus is on project LALA (Learning Analytics in Latin America) as a technician.

Ricardo Maya is a computer science engineer graduated from Escuela Superior Politécnica del Litoral (ESPOL) and a research assistant at the Information Technology Center (CTI) in Ecuador. His research is focused on education technologies, building tools that improve the students learning process and help in the decision making of the professor. He is currently working on the project LALA (Learning Analytics in Latin America) as a software developer.

Guillermo Ettlin Alonso is a Computer Engineer who has been working on ANEP as a consultant since 2014. He received his degree on 2017 from the Engineering Faculty (FING) of the University of the Republic (UDELAR), Uruguay. He is currently studying a computer master's degree by Basic Sciences Development Program (PEDECIBA), focusing on Artificial Intelligence.

Patricia Díaz is a lawyer dedicated to the analysis of technology from the perspective of Human Rights, especially interested in issues related to intellectual property and public interest, privacy and freedom of expression on the Internet, Open Data and Open Science to promote access to knowledge in the digital era. Currently, professor at Universidad de la República (Uruguay) and working on research projects related to the relationship between Copyright and Public Interest, Open Access to Scientific Research (AA), Open Educational Resources (OER), ethical aspects of Learning Analytics, Cloud Computing in Education and Educational Technologies.

ORCID

Luiz Antonio Macarini http://orcid.org/0000-0001-6127-4863
Henrique Lemos dos Santos http://orcid.org/0000-0003-0236-1291
Cristian Cechinel http://orcid.org/0000-0001-6384-409X
Xavier Ochoa http://orcid.org/0000-0002-4371-7701
Virgínia Rodés http://orcid.org/0000-0002-7229-4998
Alén Pérez Casas http://orcid.org/0000-0002-4050-2481
Ricardo Maya http://orcid.org/0000-0002-4347-202X

References

Agrawal, D. (2011). *Challenges and opportunities with big data 2011-1 (Cyber Center Technical Reports)*. West Lafayette: Purdue University. Retrieved from https://docs.lib.purdue.edu/cctech/1/.

Ali, L., Asadi, M., Gaevi, D., Jovanovi, J., & Hatala, M. (2013). Factors influencing beliefs for adoption of a learning analytics tool: An empirical study. *Computers & Education*, *62*, 130–148.

Alom, B., & Courtney, M. (2018). Educational data mining: A case study perspectives from primary to university education in Australia. *International Journal of Information Technology and Computer Science*, *2*, 1–9.

Araque, F., Roldán, C., & Salguero, A. (2009). Factors influencing university dropout rates. *Computers & Education*, *53*(3), 563–574.

Bezerra, C., Scholz, R., Adeodato, P., Lucas, T., & Ataide, I. (2016). Evasao escolar: Apli- cando mineração de dados para identificar variáveis relevantes. *Brazilian symposium on computers in education* (*Simpósio Brasileiro de Informática na Educação – SBIE*) (Vol. 27, p. 1096).

Calixto, K., Segundo, C., & de Gusmão, R. P. (2017). Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. *Brazilian symposium on computers in education* (*Simpósio Brasileiro de Informática na Educação – SBIE*) (Vol. 28, p. 1447).

Chowa, G. A., Masa, R. D., Ramos, Y., & Ansong, D. (2015). How do student and school characteristics influence youth academic achievement in Ghana? A hierarchical linear modeling of Ghana youthsave baseline data. *International Journal of Educational Development*, *45*, 129–140.

dos Santos, H. L., Cechinel, C., Nunes, J. B. C., & Ochoa, X. (2017). An initial review of learning analytics in Latin America. *2017 Twelfth Latin American Conference on Learning Technologies* (*LACLO*) (pp. 1–9). IEEE.

Duval, E. (2011). Attention please! learning analytics for visualization and recommendation. *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 9–17).

Einhardt, L., Tavares, T. A., & Cechinel, C. (2016). Moodle analytics dashboard: A learning analytics tool to visualize users interactions in Moodle.*2016 XI Latin American Conference on Learning Objects and Technology* (*LACLO*) (pp. 1–6). IEEE.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37–54.

Fernández, T., & Pereda, C. (2010). Explicar/intervenir sobre la desafiliación educativa en la enseñanza media. El Uruguay desde la Sociología VIII, 165–175.

Gibson, A., & Lang, C. (2018). The pragmatic maxim as learning analytics research method. *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 461–465).

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, *15*(3), 42–57.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18.

Hlosta, M., Zdrahal, Z., & Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 6–15).

Howard, S., & Mozejko, A. (2013). *DER-NSW evaluation: Conclusions from the 2013 data collection*. Sydney, NSW: New South Wales Department of Education and Communities. Retrieved from https://ro.uow.edu.au/cgi/viewcontent.cgi?article=1542&context=sspapers

Hübscher, R., Puntambekar, S., & Nye, A. H. (2007). Domain specific interactive data mining. *Proceedings of workshop on data mining for user modeling at the 11th international conference on user modeling* (pp. 81–90).

Hutt, S., Gardener, M., Kamentz, D., Duckworth, A. L., & D'Mello, S. K. (2018). Prospectively predicting 4-year college graduation from student applications. *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 280–289).

INEEd. (2017). *Informe sobre el estado de la educación en uruguay 2015–2016: Síntesis y desafíos*. Montevideo: Imprenta Blueprint. Retrieved from https://www.ineed.edu.uy/images/pdf/Informe-sobre-el-estado-de-la-educacion-en-Uruguay-2015-2016.pdf

Kouzes, R. T., Anderson, G. A., Elbert, S. T., Gorton, I., & Gracio, D. K. (2009). The changing paradigm of data-intensive computing. *Computer*, *42*(1), 26–34.

Macarini, L. A., Cechinel, C., Santos, H. L. D., Ochoa, X., Rodés, V., Alonso, G. E., … Díaz, P. (2019). Challenges on implementing learning analytics over countrywide K-12 data. *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 441–445). ACM.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

Martinez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39–55.

Nahum, B. (2008). Historia de educación secundaria: 1935–2008 (No. 373.9 ADMh). Administración Nacional de Educación Pública, Uruguay. Retrieved from https://eva.udelar.edu.uy/pluginfile.php/513042/mod_folder/content/0/ANEP_2008_Historia%20de%20educaci%C3%B3n%20secundaria%201935-2008.pdf?forcedownload=1

OECD. (2015). Education GPS – Uruguay – Student Performance (PISA 2015). Retrieved April 06, 2019, from http://gpseducation.oecd.org/

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Schwendimann, A., Rodriguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., … Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41.

Shea, P., & Bidjerano, T. (2014). Does online learning impede degree completion? A national study of community college students. *Computers & Education*, *75*, 103–111.

Siemens, G., & Long, P. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review*, *46*(5), 30.

Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics – A literature review. *ICTACT Journal on Soft Computing*, *5*(4), 1035–1049.

Storey, V. C., & Song, I.-Y. (2017). Big data technologies and management: What conceptual modeling can do. *Data & Knowledge Engineering*, *108*, 50–67.

Tufféry, S. (2011). *Data mining and statistics for decision making* (Vol. 2). Chichester: John Wiley & Sons.

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist*, *57*(10), 1500–1509.

Wu, J.-Y., Hsiao, Y.-C., & Nian, M.-W. (2018). Using supervised machine learning on large-scale online forums to classify course-related Facebook messages in predicting learning achievement within the personal learning environment. *Interactive Learning Environments*, 1–16. https://www.tandfonline.com/doi/full/10.1080/10494820.2018.1515085

Xu, D., & Jaggars, S. S. (2011). *Online and hybrid course enrollment and performance in Washington state community and technical Colleges*. New York, NY: Community College Research Center, Columbia University. Retrieved from https://ccrc.tc.columbia.edu/media/k2/attachments/online-hybrid-performance-washington.pdf