Use of Machine-Learning in Engineering Students' Trajectories

Martín Pratto Burgos, University of the Republic, Uruguay
Daniel Alessandrini, University of the Republic, Uruguay
Fernando Fernández, University of the Republic, Uruguay
Ximena Otegui, University of the Republic, Uruguay

The Barcelona Conference on Education 2023
Official Conference Proceedings

Abstract

Students' trajectories show the student path in the educational system from the beginning to the end of their studies. There are several statistical tools to achieve its understanding and subsequent decision-making by the institution. Each stage of the student's trajectory can be described by educational, socio-economic, demographic and cultural variables. The purpose of the research is to apply the machine learning techniques of Principal Component Analysis and k-means at the first interpretation of students' trajectories. It allows to set up clusters and prioritisation variables that organise the academic trajectory characterisation. Techniques were applied to a population of 92 Surveying students of the Engineering School at the University of the Republic (Uruguay), with admissions between 2018 and 2022. For the database processing, the statistical software R was used through RStudio, modelling five variables. In this population, data can be represented by combinations of the original variables after the Principal-Component-Analysis application. The variables that hold the highest level of importance corresponded to: Engineering School admission age and progress level determined with the obtained credits and expected credits ratio. Both variables describe the 57% of the population. On the other hand, k-means clustering has shown three groups of interest generated according to both importance variables obtained with the Principal-Component-Analysis tool. The application of machine learning techniques made it possible to plan and systematise the subsequent qualitative analysis, which included surveys and interviews.

Keywords: Machine-Learning, Students' Trajectories, Education, University

iafor
The International Academic Forum
www.iafor.org

Introduction

During the period from August to December 2022, the Engineering School Teaching and Learning Center of the University of the Republic of Uruguay carried out a monitoring of academic results and students' trajectories for the Surveying degree current curriculum. The initiative came from the Engineering School's Surveying Career Committee (SCC) based on the new curriculum implementation starting in 2023.

The SCC requires inputs to recognize areas that could benefit from improvement, leading to reduced university dropout and increased graduation rates, among other things. In this context, the general objective of the proposal was to provide the SCC an initial map of surveying students' characterisation over the last five years, as well as the main students' trajectories identification and their approach to the reasons behind them (Unidad de Enseñanza, 2023).

In the period between 2018 and 2022, 144 students were enrolled in the Surveying undergraduate degree at the Engineering School (Área Ingreso Avance Estudiantil y Rendimiento Académico [IAERA], 2023a), which corresponds to 1.4% of the total number of engineering degree programmes for the same period (IAERA, 2023b). Of the total enrolments, 92 students opted only for the Surveying undergraduate degree while the remaining 52 students opted for Surveying and other degrees from the Engineering School.

The Engineering School courses are graded on a scale of 1 to 12. The student does not pass the course when the mark obtained is 1 or 2 and must repeat the course the following year. If the mark is between 3 and 5, the student passes the course and must take a test to pass the course. Finally, a mark obtained by the student between 6 and 12 indicates that he/she does not have to take any exam because the course is taken directly (Unidad de Enseñanza, 2022).

Academic credits are gained by passing the course directly or the exam. A credit is defined as the unit of measurement of the academic work time devoted by the student to achieve the training objectives of each of the courses that make up the curriculum. Each credit is equivalent to 15 hours of student work, including hours of instruction, or equivalent activity, and hours of personal study. Each course has a certain number of credits associated with it, which requires 450 credits to obtain an engineering degree at the Engineering School (Universidad de la República, 2014).

The problem is caused by the methodology used to address a first approach to understanding student trajectories due to the numerous paths that students take at the university. A first approach to the concept of student trajectories was provided by Bourdieu (1994) through the concept of the sociological trajectory in which the agent occupies successive positions in a moving space in which he transforms.

Currently, the trajectory of a student is defined as the path students take throughout their university stay. "The trajectory is a description of the student's different positions along the curricular path, with expected times defined as normal" (Ruiz Barbot et al., 2017).

The theoretical trajectory is given by the course of the student, in accordance with the schedules stipulated by the program. On the contrary, a real trajectory is a deviation that the student makes by distancing himself from the curriculum obligations.

The first phase of comprehending students' trajectories requires gathering human and computer resources to achieve a first approximation. Determining and implementing computer and mathematical tools is crucial to begin to comprehend real student pathways.

During the fourth industrial revolution, digitalisation, artificial intelligence, the development of machine learning, and industry automation, among others, were all introduced into existence. Martnez-Ruiz (2019) argued that digital transformation was not only a catalyst for the transformation of time and space but also a crucial factor in redirecting educational challenges.

Educators are confronted with both challenges and opportunities when using artificial intelligence in education, both in their classroom practices and academic management of courses On the one hand, teachers use artificial intelligence in the classroom as a stimulating didactic tool, thus "enriching learning environments in the context of Higher Education and awakening students' interest and taste for using technologies in their future teaching practice" (Ayuso del Puerto & Gutiérrez Esteban, 2022).

Using artificial intelligence in academic management, teachers can detect different levels of university dropout risk based on course performance. Pedagogical actions can be taken based on social, geographical, cultural, and life dynamics data. Computer systems used as a tool for education "allow both to maintain and expand mass education and at the same time to develop personalised education with automatic and individual attention, both administrative and academic" (Rama, 2023).

The use of artificial intelligence should not create a gap between the use of technology and conventional methods of extracting information or resources in the classroom. To reverse the differences in technological resources, students and teachers should receive equal access in their educational institutions involving training teachers by bringing them closer to a national and international virtual environment (Ayuso del Puerto & Gutiérrez Esteban, 2022).

Virtual environments enable the use of machine-learning approaches, such as artificial neural networks, to demonstrate effectiveness in student classification. This is a significant improvement over the limitations of traditional approaches (Musso et al., 2020).

Inputs are needed for machine learning to be used. Among them, databases are relevant in students' trajectories quantitative research, taking into account variables associated with their academic performance such as achievement, university dropout, success or failure on courses, without delving into the real evolution of the admission or the students' complex situation (Guevara & Belelli, 2013).

When working with databases that require interpretation for subsequent decision-making, machine-learning techniques are often employed. Classification, regression, and clustering are all techniques used in machine learning. The paper deals with clustering techniques, specifically Principal Component Analysis (PCA) and k-means.

PCA is an unsupervised statistical method (modelling data through their relationships without comparing with previous data) that allows simplifying the complex information of sample spaces, which provide multiple variables, into a few components while preserving their information. Amat Rodrigo (2017) suggested that this technique can be very useful when used in conjunction with other statistical clustering techniques like k-means.

Like PCA, the k-means methodology corresponds to an unsupervised method. The k-means methodology “aims to divide elements into groups or clusters”, similar to PCA (López & Fernández, 2018).

University students’ trajectories are random and rarely follow a specific pattern. It will be important to find a computational alternative that allows students’ temporal descriptions to characterise them. As a result, machine learning techniques, applied to an academic database, are useful for interpreting academic trajectories.

The proposal of the research is to apply the PCA and k-means clustering techniques on a database related to the academic trajectories corresponding to students of the Surveying undergraduate degree of the Engineering School of the University of the Republic of Uruguay. Particularly, the following research questions were explored: 1. What are the two representative variables that explain the population after application of PCA technique?; 2. Are there student clusters described with the two representative variables generated from the application of the k-means technique?; 3. How can I interpret the clusters from the academic trajectories perspective?

Methodology

To process and create the database, the R statistical software (version 4.1.1) is used by Rstudio (integrated development environment) version 2021.9.1.372. The extraction of information related to student data is carried out through SQL queries to the trebol-fuentes website (Servicio Central de Informática [SeCIU], 2019) taking as a reference that the date of extraction of academic data is April 2022.

PCA and k-means techniques application require data inputs extracted from queries to the trebol-fuentes database (SeCIU, 2019) transformed from CSV format to data frames (structure for storing data sheets in R) within the programming environment in R. Each student is identified by a code provided by the trebol-fuentes database (SeCIU, 2019) related to their civil identification number or passport.

The variables to be obtained have been selected based on their relevance to give a first estimate of academic students' trajectories. They should also take into account academic progress, as well as personal and demographic information.

Numerical variables are required by PCA and k-means techniques, so gender and country of birth must be converted to numbers. The transformation is performed by using the binary-variable-generation criteria. For the student's gender, the variable 1 corresponds to male and 0 to female. For the country of birth, 1 corresponds to Uruguay and 0 to foreigners.

For the PCA and k-means techniques, “FactoMineR” (Lê et al., 2008) and “stats” packages are used respectively in the programming environment in R. Clusters’ visualisations are made with the "factoextra" package using the fviz_cluster function (Kassambara & Mundt, 2020).

Findings and Discussion

Five variables and ninety-two observations are present in the data extracted from the trebol-fuentes query (SeCIU). Observations correspond to students with two conditions simultaneously fulfilled: 1. students who opted for Surveying as the unique degree (no other chosen degree of the Engineering School), and 2. students who were admitted during the 2018-2022 period. Both conditions correspond to the 94% of the total number of Surveying students who have been admitted to the Engineering School having chosen the degree as the unique option or among others. The five variables extracted correspond to: 1. Engineering School admission age, 2. grade point average until database query date April 2022, 3. obtained credits and expected credits ratio according to the Engineering School admission year, 4. gender (male and female), 5. country of birth.

The grade point average refers to the post-course average on a scale of 1 to 12. The average may change over time as the student will receive new grades as they complete the course. The average was quantified by analysing the grades up to April 2022, when the trebol-fuentes database was queried (SeCIU, 2019).

The obtained credits and expected credits ratio refer to the ratio between the credits obtained by a student, after passing courses up to April 2022, and the expected credits according to the enrolment year to the Engineering School. The curriculum takes into account 45 credits expected by semester. Numerical values furthest from 1 indicate that the student falls behind the expected progression.

The PCA technique allows us to describe the student population using a proper combination of the original variables without losing the original insight. Figure 1 shows which of the original variables are closely related to the new dimensions (Dim), corresponding to: 1. Engineering School admission age (adm_age), 2. obtained credits and expected credits ratio (obt_exp_cred).

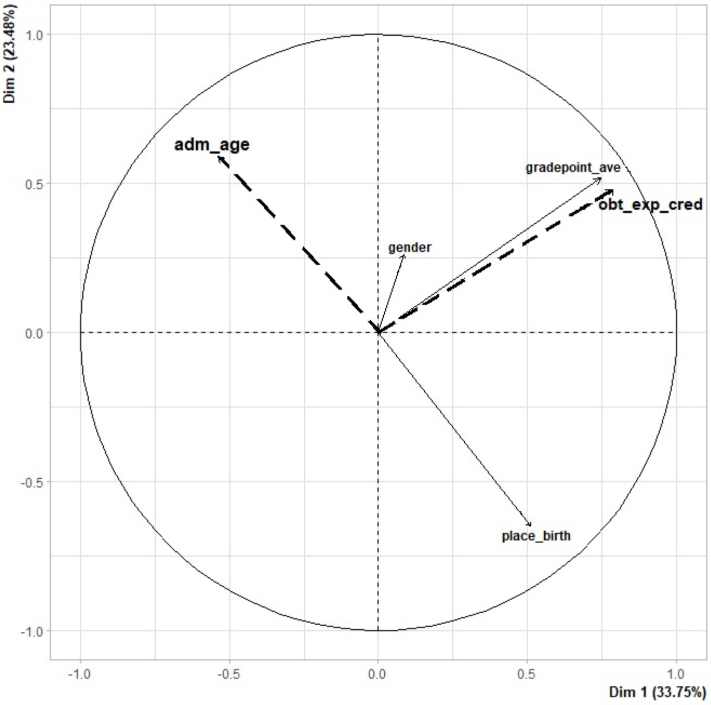

Figure 1: PCA graph of analysed variables shown through vectorial representation. Vectors in the dotted line correspond to the two best-explain population variables.

Data keeps 57.23% of the original information when it is modelled with the two crucial-variable contributions: 23.48% for the Engineering School admission age dimension (adm_age) and 33.75% for the obtained credits and expected credits ratio dimension (obt_exp_cred). The PCA technique is employed to elucidate a comprehensive explanation of the primary variables that describe data, but its interpretation is enhanced by using the k-means technique.

The joint application of PCA and k-means techniques generates three student clusters whose interpretation is centralised in the two crucial variable contributions explaining the 92 students' information dataset. The Surveying undergraduate degree has three distinct groups of students situated in well-defined quadrants with minimal overlap between them (figure 2).

Students located in quadrant 1 maintain opposite crucial variable signs: negative variable obtained credits and expected credits ratio and positive variable Engineering School admission age. Conversely, the student cluster located in quadrant 2 owns both crucial variable contributions positively. Finally, the student cluster located in quadrant 3 is formed with both crucial variable contributions negatively (figure 2).

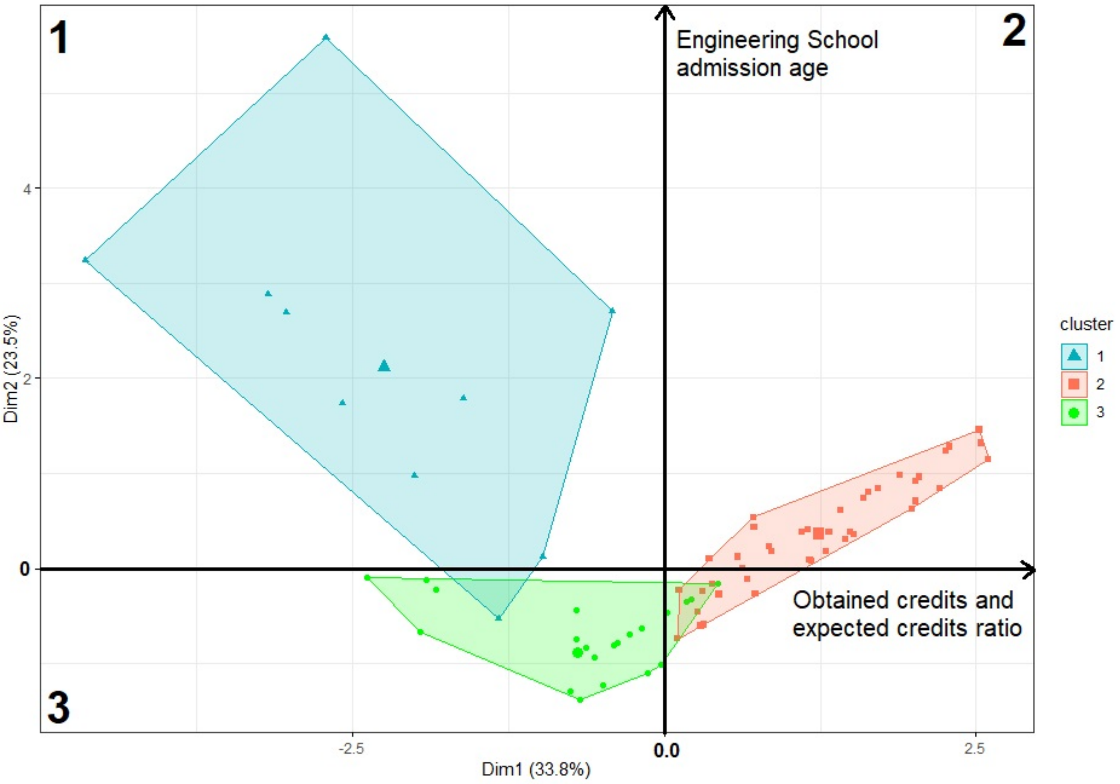

Figure 2: Two-dimensional cluster plot according to the highest contribution variables.

Unsupervised machine learning methods come from using mathematical models to create and distinguish clusters. Model interpretation is important to assign a specific trait to each group. Clusters should not overlap or they should be negligible, otherwise, it would be difficult to assign characteristics to the analysed student population.

Students admitted to the Engineering School shown in positive quadrants of figure 2 (quadrants 1 and 2) means that their admission ages are either equal to or higher than the average admission age because of the positive value axes. There is a consensus that the typical age for students to begin university is around 18 or 19. This is because they have an ideal high-school ending age corresponding to 17 years old and then they matriculate to university next year. Therefore, we interpret students located in clusters 1 and 2 as students who have an Engineering School admission age of 20 years old or older. Among them, the k-means technique divides them into two groups based on the variable obtained credits and expected credits ratio (figure 2).

Students located with negative obtained credits and expected credits ratio axis (quadrant 1 of figure 2) are associated with progress far from the ideal trajectory. This means that they are not doing well in their academic progress, lagging for the curriculum, and starting university when they are 20 years old or older (N = 10).

In contrast, students located in quadrant 2 of figure 2, with both positive crucial variables, own a good credit balance and consequently, they follow the curriculum as planned. Additionally, these cluster's students are 20 years old, or older, when they are admitted to the Engineering School (N = 41).

Finally, quadrant 3 is associated with both negative crucial variables. From one point of view, the negative Engineering School admission age axis represents the situation where students are admitted to the school at an age younger than 20 years old. From an additional point of view, the negative obtained credits and expected credits ratio axis means that their progress is not good according to the curriculum's requirements. As with quadrant 1, it implies that their progress is far from the ideal trajectory (N = 41).

Using unsupervised machine learning techniques like PCA and k-means, we can describe a population using two variables. These variables are not associated with the other three variables extracted from trebol-fuentes database (SeCIU, 2019): grade point average, gender, and country of birth. The two variables let us understand the relationship between the students' progress and the Engineering admission age through cluster creation. In such wise, cluster interpretation and data analysis after PCA and k-means techniques implementation found that Surveying students ages, lower than 20 years old, have high changes to lag for the curriculum.

Qualitative research, which involves clusters, can be enhanced with the use of other machine-learning techniques. These techniques can predict how well students will perform by considering several factors directly resulting in their potentially lagging for the degree or dropping out of university. One reason why students may take longer to complete their admitted curriculum is because they are enrolled in other engineering degrees at the same time. This can directly affect their performance in the courses they are currently pursuing (Pratto & Alessandrini, 2020).

The database information extracted includes details about students' demographics such as age, gender, university admission age, and country of birth; and their academic performance such as grade point average and obtained credits. The analysis failed to take into account other factors, resulting in an unfairly biased representation of the information. The authors Cruz et al. (2022) argued their findings in easier terms. Variations were observed in the types of data investigated through the use of machine-learning techniques. Data bias can be attributed to factors such as psychological aspects, student status in the course, and performance in certain subjects.

By clustering data according to the relationship between two variables, we gain an initial insight into the student trajectories. Results from PCA and k-means techniques application provide valuable qualitative perspectives on students in each cluster, helping us understand the first impact at university. We can enhance a detailed description of the individual groups by including additional characteristics of the population obtained by sampling each cluster. Each one can be arbitrarily sampled to collect information from each student by interviews. Finally, it aids in gathering social information reducing data bias restricted by the mathematical models.

Conclusion

Applying the PCA technique allows one to gain an initially clearer understanding of students' trajectories in the Surveying field. It focuses on two crucial variable contributions extracted from the trebol-fuentes database (SeCIU, 2019): Engineering School admission age and obtained credits and expected credits ratio. These variables can model the data while maintaining the value of 57.23% on the original information. Additionally, the k-means technique, when used with PCA, enables students to be classified into three distinct clusters depending on their positive or negative variables and placement within the graph quadrant.

The initial group, located in quadrant 1, gathers an opposing correlation between the two crucial variable contributions. Primarily, the negative obtained credits and expected credits ratio axis suggested students who have a lower number of credits compared to their ideal trajectory. Secondly, this students’ cluster includes those who were enrolled in the Engineering School older than the admission age average, provided by the positive Engineering School admission age axis. The association between both variables suggests that students admitted to university at the age of 20 years old, or older, demonstrate unsatisfactory academic development about the theoretical trajectory given by the curriculum (N = 10).

Quadrant 2 includes the second cluster, comprising students who display both positive crucial variables, such as Engineering School admission age and obtained credits and expected credits ratio. A satisfactory student's academic progression at the university admission age of 20, or older, is associated with both positive variables (N = 41).

Finally, the third group, located in quadrant 3, is composed of students with the negative crucial variables. Insufficient academic progress in the Surveying undergraduate degree is a consequence of admission to Engineering School at the age of less than 20 years (N = 41).

Students’ trajectories can be analysed by using unsupervised machine learning methods like PCA and k-means. Both methods initially contribute to comprehend the behaviour of students in their university degree.

Using academic information as a source, clustering helps us to continue the data analysis using predictive machine-learning techniques and also organising a sampling. It means that we can gather social information about student's experiences by talking to them. It avoids the bias caused by emotions or other psychological factors that might not be considered in academic databases.

References

Amat Rodrigo, J. (2017). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE.* Attribution 4.0 International (CC BY 4.0). https://www.cienciadedatos.net/documentos/35_principal_component_analysis

Área Ingreso Avance Estudiantil y Rendimiento Académico - IAERA (2023a). *Informe de Indicadores de Seguimiento del Plan de Estudios 2018-2022*. Unidad de Enseñanza, Facultad de Ingeniería, Universidad de la República, Uruguay.

Área Ingreso Avance Estudiantil y Rendimiento Académico - IAERA (2023b). *Informe de Avance Estudiantil de las Carrera de Grado Ingenieriles 1997-2022*. Unidad de Enseñanza, Facultad de Ingeniería, Universidad de la República, Uruguay.

Ayuso del Puerto, D., & Gutiérrez Esteban, P. (2022). La Inteligencia Artificial como recurso educativo durante la formación inicial del profesorado. *RIED-Revista Iberoamericana De Educación a Distancia*, *25*(2), 347–362. https://doi.org/10.5944/ried.25.2.32332

Bourdieu, P. (1994). *Razones prácticas sobre la teoría de la acción.* Traducción: Thomas Kauf. Editor digital: diegoan ePub base r1.2

Colors in R. R Charts. [Online] https://r-charts.com/es/colores/ Consultado 20 de diciembre de 2022.

Cruz, E., González, M., Rangel Ortiz, J. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento y a la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico*, 13, 77-87. 10.33412/pri.v13.1.3039.

Guevara, H., and Belelli, S. (2013). Las trayectorias académicas: dimensiones personales de una trayectoria estudiantil. Testimonio de un actor. *RevIISE - Revista De Ciencias Sociales Y Humanas*, 4(4), 45-56.

Kassambara, A. and Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. https://CRAN.R-project.org/package=factoextra

Lê S., Josse J., Husson F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1), 1-18. 10.18637/jss.v025.i01

López, D. & Fernández, A. (2018). Aplicación en los medios de prensa de un agrupamiento k-means (clustering k-means). *Revista Chilena de Economía y Sociedad*, 12(1), 26-48.

Martínez-Ruiz, X. (2019). La industria 4.0. y las pedagogías digitales: aporías e implicaciones para la educación superior. *Innovación Educativa*, 19(79), 7-12. https://bit.ly/3caSiyD

Musso, Mariel F., Hernández, Carlos Felipe Rodríguez, Cascallar, Eduardo C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. *Higher Education* 80 (5), 875–894. https://doi.org/10.1007/s10734-020-00520-7

Pratto, M., Alessandrini, D. (2020). Egresos con inscripciones multicarreras de Ingeniería. *InterCambios*. Dilemas y transiciones de la Educación Superior, 7(1), 91-97. https://ojs.intercambios.cse.udelar.edu.uy/index.php/ic/search/authors/view?firstName=Mart%C3%ADn&middleName=&lastName=Pratto%20Burgos&affiliation=Autor&country=UY

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rama, Claudio (2023). *La educación personalizada y la inteligencia artificial*. Columna de Opinión, Grupo R Multimedio. [Online] https://grupormultimedio.com/la-educacion-personalizada-y-la-inteligencia-artificial-id74223/ Consultado, 26 de mayo de 2023.

Rdocumentation. fviz_cluster: Visualize Clustering Results. factoextra (version 1.0.7). [Online] https://www.rdocumentation.org/packages/factoextra/versions/1.0.7/topics/fviz_cluster Consultado 20 de diciembre de 2022

Ruiz Barbot, M., Fachinetti, V., Barceló, J., Romero, P. (2017). *Los estudiantes universitarios, trayectos de formación. Sujetos Contemporáneos, aprendizaje y comunicación.* Jornadas de Investigación en Educación Superior, Montevideo, Uruguay.

Servicio Central de Informática - SeCIU (2019). Proyecto Trébol. Recuperado de la base de datos de trebol_fuentes v1.0.1. Universidad de la República, Uruguay.

Unidad de Enseñanza (2022). *Desempeño estudiantil en Unidades Curriculares de primer año.* Facultad de Ingeniería, UdelaR. Montevideo, Uruguay.

Unidad de Enseñanza (2023). *Seguimiento de trayectorias estudiantiles de la Carrera de Agrimensura del Plan de estudios 1997, 2013-2022.* Facultad de Ingeniería, UdelaR. Montevideo, Uruguay.

Universidad de la República (2014). *Ordenanza de estudios de grado y otros programas de formación terciaria: normativa y pautas institucionales relacionadas.* Comisión Sectorial de Enseñanza, Unidad Académica. Temas de enseñanza (1). Montevideo, Uruguay.

Contact email: martinpburgos@gmail.com