



Education in the Knowledge Society

journal homepage <http://revistas.usal.es/index.php/eks/>

Predicting university dropout using machine learning: Applications based on multidisciplinary databases

Prediciendo el abandono universitario utilizando técnicas de aprendizaje automático: Aplicaciones sobre bases de datos multidisciplinarias

Mauricio Barramuño-Medina^a, Pablo Valdés-Badilla^b, José Luis Saavedra-Lucero^c, Germán Gálvez-García^d, Relmu Gedda-Muñoz^{e(*)}^a Kinesiology program, Faculty of Health Sciences, Universidad Autónoma de Chile, Temuco, Chile, <https://orcid.org/0000-0002-0961-2869> mauricio.barramuno@uautonoma.cl^b Department of Physical Activity Sciences, Faculty of Education Science, Universidad Católica del Maule, Talca, Chile; and Sport Trainer program, School of Education, Universidad Viña del Mar, Viña del Mar, Chile<https://orcid.org/0000-0002-3948-8280> pvaldes@ucm.cl^c Academic Vice-rectory, Universidad Autónoma de Chile, Temuco, Chile<https://orcid.org/0000-0001-8120-7510> jose.saavedra@uautonoma.cl^d Departamento de Psicología Experimental, Psicobiología y Metodología de Ciencias del Comportamiento, Universidad de Salamanca, Salamanca, España; and Departamento de Psicología, Universidad de La Frontera, Temuco, Chile<https://orcid.org/0000-0001-7390-4180> germangalvezgarcia@usal.es^e Sociology Department, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, España; and Academic Vice-rectory, Universidad Autónoma de Chile, Chile <https://orcid.org/0000-0002-9747-3220> relmu.gedda@uautonoma.cl

ARTICLE INFO

Keywords

student dropout; students; higher education; prediction; machine learning; data mining.

ABSTRACT

Student dropout constitutes a challenge for higher education institutions due to, among other reasons, the diversity of students' social, educational, and economic backgrounds, as well as the disciplines and academic degrees in which they enrol, resulting in a wide variability in educational trajectories. The study aimed to develop a predictive model of university dropout among students enrolled in various programs, utilising machine learning. An online questionnaire and academic records were used to obtain ten variables previously identified as relevant to the studied phenomenon, and classifiers were trained through supervised machine learning procedures. The sample consisted of 946 students from seven different academic disciplines at a university in Chile. The primary outcomes indicate that a model based on the RUSBoosted Trees classifier presented the best performance in predicting dropout in higher education, successfully recognising 90% of dropout students. Several previous studies have predicted student dropout using machine learning, but typically on samples of students belonging to the same programmes. This study presents a predictive model that identifies future dropouts even under multidisciplinary conditions within their original programmes.

RESUMEN

Palabras clave

deserción estudiantil; estudiantes; educación superior; predicción; aprendizaje automático; minería de datos.

La deserción estudiantil constituye un desafío para las instituciones de Educación Superior debido, entre otras razones, a la diversidad de los estudiantes en cuanto a su origen social, educativo y económico, así como a las disciplinas y grados académicos en los que se matriculan, produciendo una amplia variabilidad en las trayectorias educativas. El objetivo de este estudio fue desarrollar un modelo predictivo para el abandono universitario en estudiantes pertenecientes a diferentes titulaciones, basado en machine learning. Se utilizaron un cuestionario *online* y los expedientes académicos del estudiantado para obtener diez variables previamente identificadas como relevantes para el fenómeno estudiado y se entrenaron clasificadores mediante procedimientos de

(*) Corresponding author / Autor de correspondencia.

aprendizaje automático supervisado. La muestra incluyó 946 estudiantes pertenecientes a siete áreas de conocimiento en una universidad de Chile. Los principales resultados indican que un modelo basado en el clasificador RUSboosted Trees presentó el mejor desempeño para la predicción de la deserción en Educación Superior, reconociendo exitosamente al 90% de las y los estudiantes desertores. Varios estudios previos han enfrentado la tarea de proporcionar herramientas para la predicción de la deserción estudiantil utilizando aprendizaje automático, pero generalmente sobre muestras de estudiantes pertenecientes a las mismas titulaciones o áreas del conocimiento. Este estudio proporciona un modelo predictivo que identifica a los futuros estudiantes desertores, incluso en condiciones de multidisciplinariedad dentro de sus programas de origen.

1. INTRODUCTION

Higher education (HE) systems worldwide have experienced a progressive increase in enrolment, diversification of institutions, transformation of their structures, as well as a series of pedagogical and curricular adjustments to adapt to a highly dynamic environment (OECD, 2020, 2023). In Chile, HE has faced a similar situation in recent years, with the variability of university entrance profiles presenting the most significant challenges for institutions (Gedda-Muñoz & Carrasco-Bahamonde, 2023; Santelices et al., 2021; Teixeira et al., 2017). With the increase in coverage and access to HE, associated phenomena have emerged that require evaluation, such as student dropout and the implementation of strategies to prevent HE students from dropping out (Aulck & West, 2017; Braxton, 2019; Gedda-Muñoz & Villagrán, 2020; Larroucau, 2016).

Dropout in HE refers to the process of leaving a study program prematurely, voluntarily, or forcibly, which prevents the student from obtaining a degree (Braxton, 2019; Tinto, 1975, 2022), considering a lapse of at least three years to rule out the possibility of the student returning to HE (SIES, 2019).

University dropout entails family and social costs that justify the adoption of public and institutional policies aimed at reducing abandonment rates and increasing students' graduation prospects. Tinto (1975, 2022) is one of the scholars who initially contributed to understanding student attrition, conceptualising it as a process of abandonment influenced by internal and external circumstances. Moreover, other studies have meticulously examined university dropout to comprehend and address this phenomenon (Aulck & West, 2017; Braxton, 2019; Donoso & Schiefelbein, 2007; Himmel, 2002). Analysing the different approaches to dropout, three groups of explanatory factors emerge: sociodemographic, individual, and institutional. Sociodemographic factors, such as students' family socioeconomic status, country of origin, residential area (urban or rural), and parental education level, have been described in previous research (Barramuño et al., 2022; Heredia et al., 2015; Kuz & Morales, 2023; Palacios et al., 2021; Sani et al., 2020; Siri, 2015). Concerning individual factors, gender and age, students' high school academic backgrounds, and scores on university admission tests, as well as university academic performance measured by grade point average, credits, or passed and failed courses, have been identified (Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Berens et al., 2018; Eegdeeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Palacios et al., 2021; Sani et al., 2020; Siri, 2015). Additionally, factors related to student engagement, academic self-concept, motivation, and study habits have been explored (Cancino-Marchant et al., 2022; Christou et al., 2023; Eegdeeman, Cornelisz, van Klaveren et al., 2022; Gedda-Muñoz & Villagrán, 2020; Palacios et al., 2021; Parra-Sánchez et al., 2023).

Regarding institutional factors, previous studies on university dropout have primarily described variables such as funding modalities available to students, including scholarships or loans, the existence of special access, levelling, or academic support programs, degree duration, and curriculum completion requirements, among others (Barramuño et al., 2022; Beer & Lawson, 2016; Christou et al., 2023; Heredia et al., 2015; Himmel, 2002; Kemper et al., 2020; Kuz & Morales, 2023; Larroucau, 2016; Palacios et al., 2021; Sani et al., 2020; Santelices et al., 2021; Siri, 2015). However, it is important to note that none of these factors, whether sociodemographic, individual, or institutional, operate in isolation. Various pieces of evidence indicate the influence and interaction among these different levels, highlighting the complexity and interdependence of students' educational trajectories (Bourdieu & Passeron, 2019; Reay, 2004, 2022; Soto-Hernández, 2016). Research on university dropout has sought to provide evidence to reduce educational attrition and promote successful educational trajectories, as individuals who do not complete their studies diminish their social mobility prospects and miss out on acquiring key competencies for future employability (Kuz & Morales, 2023; Larroucau, 2016), leading to negative social and economic impacts for governments, HE institutions, and students' families (Aulck & West, 2017; Beer & Lawson, 2016).

Furthermore, studies employing machine learning to predict student attrition dynamics have highlighted Decision Tree (Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Christou et al., 2023; Kemper et al., 2020; Kuz & Morales, 2023; Limsathitwong et al., 2018; Miranda & Guzmán, 2017; Ortiz-Lozano et al., 2018; Palacios

et al., 2021; Sani et al., 2020; Wan Yaacob et al., 2020), Artificial Neural Networks (Kuz & Morales, 2023; Miranda & Guzmán, 2017; Sani et al., 2020; Siri, 2015; Wan Yaacob et al., 2020), Support Vector Machine (Christou et al., 2023; Eegdeman, Cornelisz, Meeter et al., 2022; Liang et al., 2016; Palacios et al., 2021) and K-Nearest Neighbour (Barramuño et al., 2022; Palacios et al., 2021; Wan Yaacob et al., 2020) as the most commonly used classifiers with higher effectiveness rates (Fahd et al., 2021; Parra-Sánchez et al., 2023). They have also reported that variables related to students' academic performance, assessed by grade point average and semester grades, as well as the number of failed courses or exams, exhibit greater predictive capacity than other sociodemographic (socioeconomic status, parental education level) and institutional variables (funding type, special access and academic support programs). In this regard, academic performance, measured by students' grades, contributes to greater predictive capacity as it reflects the outcome of multiple interacting factors, including personal, relational, social, and institutional factors (Garbanzo Vargas, 2012; Zwick, 2012).

In Chile, public policies have promoted student retention in HE, which has made it possible to implement accompaniment strategies that foster the permanence of students in the educational system (Aljohani, 2016; Pitman et al., 2014). Although various actions and programs that aim at reducing dropouts have been introduced, the outcomes are far from significant: only 3% in the last ten years for Chilean first-year HE students (SIES, 2023). The first-year dropout rate for Chilean university students is around 20% (SIES, 2023), a percentage that increases to 31% in the second year of study (CNED, 2022) and reaches levels between 36% and 46% at the end of the third year (Himmel, 2002; OFD-CIAE, 2021). These facts reflect the need to propose new information strategies and remedial programs to help reduce dropout rates among university students.

The inclusion of new technologies, especially the use of strategies based on computational tools such as machine learning, has emerged, which is used for the recognition of information patterns using large databases, aiming to generate predictive models (Cuji Chacha et al., 2020; Dutt, 2015; Eegdeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kuz & Morales, 2023), allowing the identification of risk factors, which serve as input to undertake anticipatory actions. In addition, these databases are constantly updated, which provides validity and reliability to the information, making institutional data repositories beneficial sources for models based on machine learning. Previous studies using machine learning have reported predictive models of university dropout with accuracy levels ranging from 70% to 98% (Parra-Sánchez et al., 2023; Sani et al., 2020). These results vary depending on the type of classifier, the procedures employed, and the variables used. On the other hand, the predictive models proposed by previous studies have the limitation of having homogeneous samples, i.e., set in degrees belonging to the same area of knowledge, such as computer science and informatics (Abu-Oda & El-Halees, 2015; Limsathitwong et al., 2018; Wan Yaacob et al., 2020), engineering (Miranda & Guzmán, 2017; Ortiz-Lozano et al., 2018), applied sciences (Berens et al., 2018), or health sciences (Barramuño et al., 2022; Siri, 2015). Data homogeneity hinders the transferability of models to other programs or institutions, as it is uncertain whether the results obtained in an engineering degree can be reliably extrapolated to architecture or design.

On the other hand, the studies carried out present the accuracy or precision of their models as an indicator of quality and predictive capability, but only sometimes emphasise other equally essential measures such as F-score or sensitivity (Eegdeman, Cornelisz, Meeter et al., 2022). To interpret the effectiveness of a predictive model, special attention must be paid to the nature of the phenomenon being explained and to the metrics that best achieve this task.

In this context, proposing predictive models that are representative and applicable to students from different degrees and faculties could be an opportunity for academics and management staff to make more accurate and contextualised decisions about students, helping them in their curricular progress and educational trajectory. The following study questions were raised: (i) What is the most effective machine learning classifier for identifying higher education students at risk of dropout using multidisciplinary databases? Once the classifier is identified, it could be established: (ii) What is the most effective specific predictive model for identifying higher education students at risk of dropout using multidisciplinary databases? Accordingly, this study aims to develop a multidisciplinary predictive model of dropout risk among higher education students using machine learning.

2. MATERIALS AND METHODS

2.1. Participants

The sample was selected under an accidental nonprobability criterion and initially consisted of 1,374 students (73.9% female and 26.1% male) from a private university in Chile. An accidental non-probabilistic

criterion was chosen because there was a need to complement existing databases. Although the students' university academic records were available, it was recognised that certain variables relevant to the dropout phenomenon, such as the educational level of the mother and father, were absent. Consequently, a voluntary survey was conducted among students. Subsequently, upon merging both datasets (university academic records and the survey data), following inclusion criteria were applied: (i) being 18 years old or older without distinction of gender; (ii) being a regular student with an active enrolment, along with the following exclusion criteria: (i) students in the graduation process; and (ii) students who answered the instrument partly. Finally, after data processing and filtering, the sample consisted of 946 students (700 females and 246 males) with a mean age of 23.2 ± 3.2 years. In Tables 2 and 3, you can find sample characterisation.

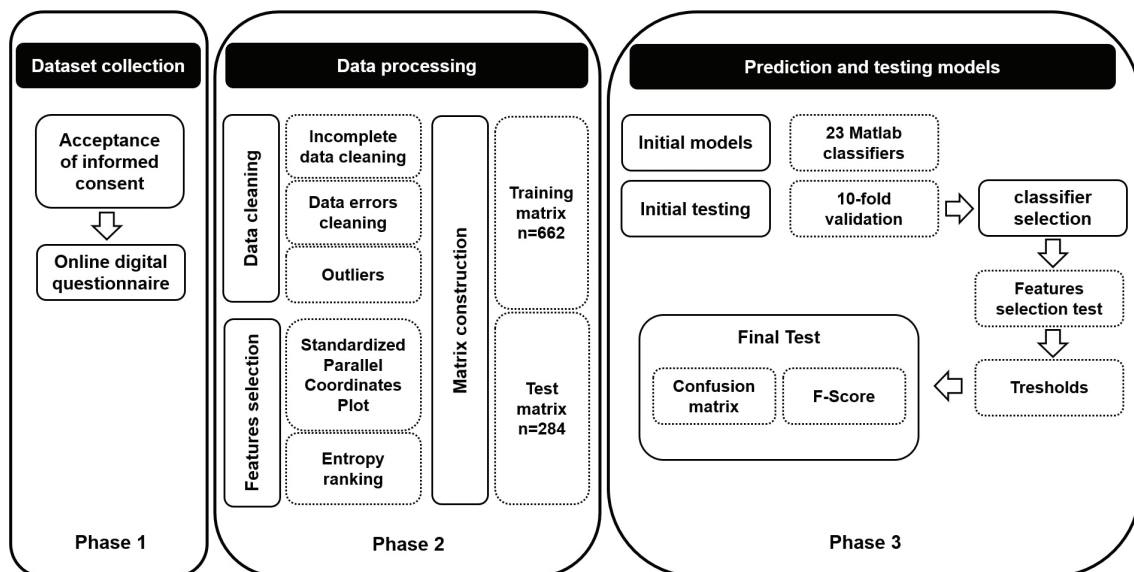
The assessed students were informed of the study's aim, scope, and implications, and provided informed consent that authorises the use of the information for scientific purposes. The research protocol was developed in accordance with the Declaration of Helsinki and was approved by the Scientific Ethical Committee (n°120-18) of the Universidad Autónoma de Chile.

2.2. Dataset collection

The dataset collection process consisted of a voluntary and self-administered online questionnaire, along with explicit informed consent, and the capture of academic and sociodemographic information from the academic records databases.

The stages involved in performing the collection, data processing, analysis, prediction, and testing of models are illustrated in Figure 1.

Figure 1. Stages of data collection and processing procedures, analysis, prediction and testing models.



2.3. Variables used in the study

The variable selection was based on the literature review findings, paying particular attention to ensure that they were similar to those recorded in the universities' administrative databases, following the recommendations of studies employing machine learning in the university context (Berens et al., 2018; Fahd et al., 2021; Kemper et al., 2020; Kuz & Morales, 2023; Parra-Sánchez et al., 2023)—this proposal aimed to develop a predictive model using machine learning that other institutions could utilise. The highlighted variables were the mother's educational level, father's educational level, age, gender, school origin, years to initiate HE (after secondary school), financing type, number of subjects failed, grade point average, and number of subjects enrolled per year.

The variables regarding the educational level of the mother and father were obtained through the online questionnaire, while the age, gender, type of school, years to enter university, tuition financing type, number of

subjects failed, grade point average, number of subjects enrolled per year, and student dropout variables were obtained from the university. The academic record is compiled from information gathered during the diagnostic and induction processes at the beginning of the first year of study (entrance to the university). It is periodically complemented according to the curricular and academic progress of the students. In Table 1 are the variables used in this study, along with previous research that has recommended their use.

Table 1. Variables used in this study and in previous research.

Variable	Previous studies
Grade point average	(Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Berens et al., 2018; Christou et al., 2023; Eegdeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Kuz & Morales, 2023; Limsathitwong et al., 2018; Miranda & Guzmán, 2017; Ortiz-Lozano et al., 2018; Palacios et al., 2021; Sani et al., 2020; Siri, 2015; Wan Yaacob et al., 2020)
Gender	(Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Berens et al., 2018; Eegdeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Kuz & Morales, 2023; Ortiz-Lozano et al., 2018; Palacios et al., 2021; Sani et al., 2020; Siri, 2015; Wan Yaacob et al., 2020)
Age	(Barramuño et al., 2022; Berens et al., 2018; Eegdeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Kuz & Morales, 2023; Palacios et al., 2021; Sani et al., 2020; Siri, 2015)
Tuition funding	(Barramuño et al., 2022; Heredia et al., 2015; Kuz & Morales, 2023; Miranda & Guzmán, 2017; Sani et al., 2020; Siri, 2015)
Number of subjects failed	(Barramuño et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Palacios et al., 2021)
Mother's education level	(Barramuño et al., 2022; Heredia et al., 2015; Kuz & Morales, 2023; Siri, 2015)
Father's education level	(Barramuño et al., 2022; Heredia et al., 2015; Kuz & Morales, 2023; Siri, 2015)
Secondary School	(Barramuño et al., 2022; Miranda & Guzmán, 2017; Ortiz-Lozano et al., 2018; Palacios et al., 2021)
Number of subjects enrolled per year	(Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Heredia et al., 2015)
Years to university entrance	(Berens et al., 2018; Christou et al., 2023; Miranda & Guzmán, 2017)

To estimate student dropout, students enrolled in the same academic period as those participating in this study and who did not enrol again in the following year were used as a parameter, excluding those who had begun their graduation process.

Below is a characterisation of the study variables according to the faculty to which the students belong. Table 2 shows the percentage distribution of the categorical variables. Table 3 shows the mean and standard deviation of the numerical variables, differentiated by academic status.

Table 2. Characterisation of the sample on categorical variables.

Variable	Category	% ARC ^a (n=21)	% BA (n=84)	% EDU (n=102)	% ENG (n=60)	% HEA (n=448)	% LAW (n=53)	% SOC (n=178)	Overall (n=946)
Academic status	Dropouts (n=36)	19.0	4.8	1.0	3.3	4.7	1.9	1.7	3.8
	No dropouts (n=910)	81.0	95.2	99.0	96.7	95.3	98.1	98.3	96.2
Gender	Female (n=700)	38.1	64.3	75.5	40.0	83.9	62.3	71.9	74.0
	Male (n=246)	61.9	35.7	24.5	60.0	16.1	37.7	28.1	26.0
Mother's educational level	Up to complete high school (n=589)	81.0	66.7	82.4	65.0	56.7	47.1	64.1	62.3
	Up to technical degree (n=216)	14.3	20.2	8.8	21.7	25.9	32.0	23.1	22.8
	Up to complete university degree (n=128)	4.8	11.9	8.8	10.0	15.6	18.9	12.3	13.5
	Postgraduate (n=8)	-	-	-	1.7	1.1	1.9	0.6	0.8

(continued)

Table 2. Characterisation of the sample on categorical variables. (*continued*)

Variable	Category	% ARC ^a (n=21)	% BA (n=84)	% EDU (n=102)	% ENG (n=60)	% HEA (n=448)	% LAW (n=53)	% SOC (n=178)	Overall (n=946)
Father's educational level	Up to complete high school (n=547)	85.8	63.0	76.4	55.0	51.4	49.1	61.3	57.8
	Up to technical degree (n=178)	9.5	19.1	11.8	18.3	22.1	17.0	16.3	18.8
	Up to complete university degree (n=144)	4.8	10.7	5.9	11.7	18.6	26.4	13.5	15.2
	Postgraduate (n=17)	-	-	1.0	3.3	1.8	1.9	2.8	1.8
Type of school	Public (n=252)	42.9	21.4	42.2	25.0	22.8	34.0	26.4	26.6
	Charter (n=663)	57.1	77.4	56.9	71.7	72.3	64.2	71.3	70.1
	Private (n=24)	-	1.2	-	1.7	4.0	1.9	1.7	2.5
	Delegated administration (n=7)	-	-	1.0	1.7	0.9	-	0.6	0.7
Tuition funding type	No gratuity (n=280)	14.3	20.2	20.6	15.0	39.1	17.0	25.4	29.5
	With gratuity (n=666)	85.7	79.8	79.4	85.0	60.9	83.0	74.6	70.5

^a ARC= Architecture, Construction and Environment; BA= Business and administration; EDU= Education; ENG= Engineering; HEA= Health Sciences; LAW=Law; SOC= Social Sciences and Humanities.

Table 3. Characterisation of the sample on numeric variables.

Variable	Category	ARC ^a (n=21)	BA (n=84)	EDU (n=102)	ENG (n=60)	HEA (n=448)	LAW (n=53)	SOC (n=178)	Overall Mean	Overall Sd
Age (years)	Dropout (n=36)	26.5	23.0	24.0	25.5	24.6	22.0	22.7	24.4	3.0
	No-dropout (n=910)	22.2	22.3	24.0	21.7	23.4	23.1	22.9	23.1	3.2
Years to initiate higher education	Dropout (n=36)	1.8	2.8	3.0	3.5	3.0	2.0	1.3	2.7	1.8
	No-dropout (n=910)	2.1	2.3	3.4	2.1	2.8	2.7	2.8	2.7	2.6
Number of subjects enrolled per year	Dropout (n=36)	8.8	9.3	11.0	10.0	7.6	12.0	9.7	8.6	2.7
	No-dropout (n=910)	11.7	11.1	11.8	9.9	9.2	10.9	10.3	10.1	3.4
Number of subjects failed	Dropout (n=36)	4.5	5.0	1.0	2.5	1.6	2.0	1.3	2.4	2.6
	No-dropout (n=910)	1.2	0.7	0.4	1.4	0.3	0.8	0.2	0.5	1.1
Grade point average ^b	Dropout (n=36)	3.3	3.1	4.6	4.5	4.3	4.0	3.9	4.0	1.2
	No-dropout (n=910)	5.0	5.1	5.7	4.6	5.4	5.1	5.6	5.4	0.7

Note: All values are arithmetic means, except Standard deviation.

^a ARC= Architecture, Construction and Environment; BA= Business and administration; EDU= Education; ENG= Engineering; HEA= Health Sciences; LAW=Law; SOC= Social Sciences and Humanities; Sd= Standard deviation.

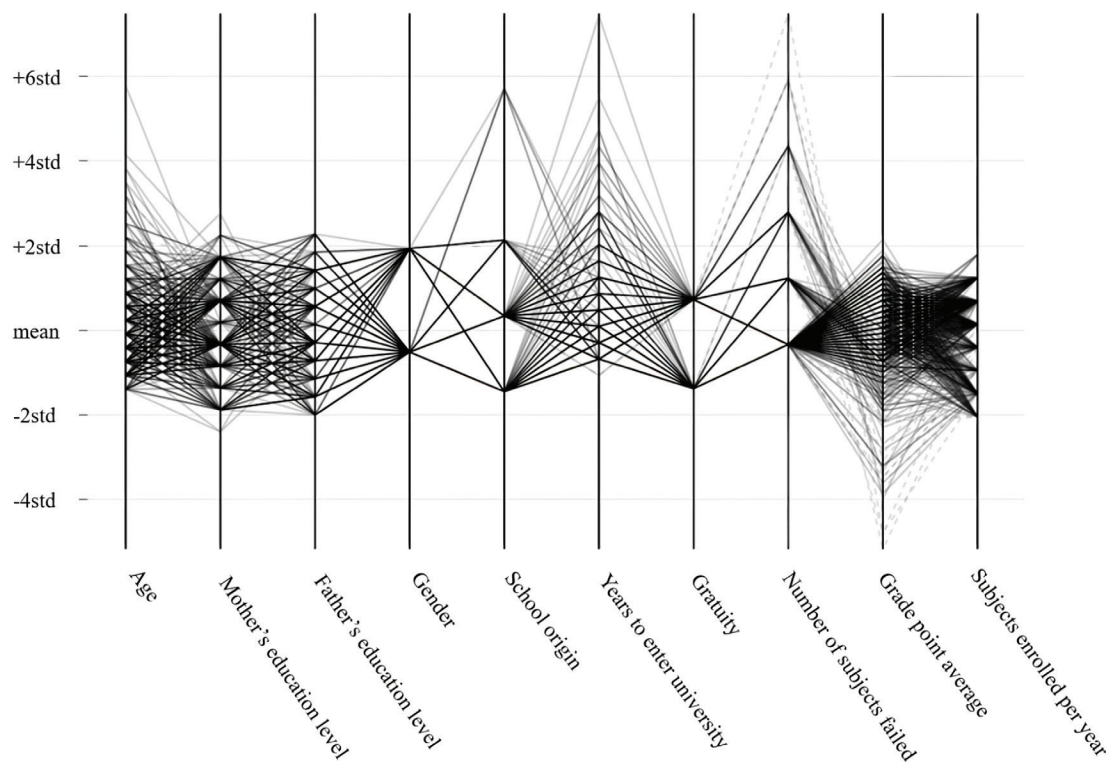
^b Grade scale from 1.0 to 7.0

2.4. Dataset processing and selection

A matrix of 1,374 students was initially obtained. All variables studied were converted into numerical values. Records of first-year students ($n = 412$) were excluded because of the high heterogeneity of the information, and extreme data ($n = 16$) were removed, resulting in a matrix of 946 students.

Subsequently, the matrix was randomly divided into two parts: (i) classifier training matrix, which corresponds to 70% of the data ($n = 662$), and (ii) test matrix, which represents 30% of the data ($n = 284$). The partitioning of data into training and testing sets in machine learning is crucial for objectively evaluating the model's performance on unseen data. It helps to avoid overfitting and understand the model's actual predictive capacity. The specific proportion of testing data, such as the 30% utilised in this study, is determined by the size of the dataset to ensure a precise assessment. Previous studies (Berens et al., 2018; Eegdeman, Cornelisz, Meeter et al., 2022; Kemper et al., 2020; Parra-Sánchez et al., 2023) report that a test set of 30% of the data is typically used, highlighting a broad consensus on this aspect. Figure 2 presents a parallel coordinate plot to illustrate the distribution of dropout and non-dropout cases across the study variables.

Figure 2. Standardised Parallel Coordinates Plot for the variables associated with dropout.



Note: Standardisation plots set the mean of each predictor to zero and scaled the predictors by their standard deviations. Solid line: non-dropping student; Segmented line: student dropout.

Then, through WEKA software and the InfoGainAttributeEval function (Frank et al., 2017), an entropy ranking of the variables was obtained (Table 4). According to Choudhary et al. (2014), entropy is commonly used in information theory measurement. This method measures the significance of an attribute by calculating the information gain concerning the target class. Based on the returned ranking of attributes, the dataset is modified. Those attributes that contribute more information will have a higher gain value and can be selected.

In contrast, those who add little information will have a lower score and may be removed (Gnanambal et al., 2018). In this regard, interpreting machine learning models can be challenging due to the inherent complexity of specific algorithms, making it difficult to explain the individual contribution of variables and their significant inclusion in the results. This entropy analysis helps identify the most informative features for classification, thereby contributing to the model's effectiveness. In simple terms, variables with higher entropy are more explanatory of the model's performance and predictive capacity.

Table 4. Entropy ranking of the variables used in the study.

Variable	Operative definition	Ranking
Grade point average	Grade scale from 1.0 to 7.0	1
Number of subjects failed	Number of subjects failed in last year	2
Mother's education level	a. Unknow, b. up to complete high school, c. up to technical degree, d. up to complete university degree, e. postgraduate, f. military academic training	3
Number of subjects enrolled per year	Number of subjects taken in the year	3
Gender	Male - Female	3
Father's education level	a. Unknow, b. up to complete high school, c. up to technical degree, d. up to complete university degree, e. postgraduate, f. military academic training	3
School	a. Public, b. Charter, c. Private	3
Tuition funding	Gratuity: Yes – No	3
Years to university entrance	Years elapsed between the end of high school and university entry	3
Age	Age in completed years	3

As can be seen, entropy analysis revealed that students' grades during the semester and the number of subjects failed in the previous academic period are the variables that most significantly contribute to the model's predictive capacity. On the other hand, the variables ranked in position three are grouped as attributes that contribute similarly to the model's performance.

2.5. Prediction models

The machine learning application of the Matlab® software (R2012b version, USA) was used to create prediction models using the training matrix obtained in the previous process. Twenty-three statistical classifiers were trained with different hyperparameters, all based on supervised machine learning.

The Matlab® "Classification Learner" application provides an interactive interface for training classification models without programming, allowing users to explore and analyse data intuitively. Upon opening the application, users can load their datasets from various sources, including CSV and Excel files. Through this tool, users can select relevant features and choose from a variety of available classification algorithms, including SVM and decision trees. The application automatically adjusts the model's hyperparameters during the training process and provides detailed performance metrics to evaluate the model's quality. Additionally, it offers visual tools for analysing the model's performance, such as confusion matrices and ROC curves. Once a satisfactory model is obtained, users can automatically generate the corresponding MATLAB code to replicate the training and classification process, as well as export the trained model for use in other applications. In this regard, to facilitate its usage, we have developed open-access code for applied researchers. This code includes the trained classification model and an example. The output represents the classifier's result for each student, coded as 0 for zero dropout and 1 for non-dropout. The code has been uploaded to the following repository.

The procedures based on supervised machine learning can be of two types: regression, which seek to develop a model that predicts the behaviour of a continuous variable based on one or more related variables, focusing on finding the solution to reduce the distance between the real data and the model curve (Barramuño et al., 2022; Nasteski, 2017); or classification, which seeks to determine the conditions met by the cases of a database to be correctly classified according to a target variable, generating the ability to predict the future behaviour of new cases that comply with the rules of the model (Tomasevic et al., 2020). Supervised learning instructs the machine to learn from our data by specifying a target variable. This reduces the machine's task to identifying patterns in the input data to predict the target variable (Harrington, 2012).

Among the classification algorithms for the phenomenon under study, support vector machines (Liang et al., 2016; Tomasevic et al., 2020), neural networks (Márquez-Vera et al., 2016; Wan Yaacob et al., 2020), and decision trees are notable. The latter corresponds to a classifier that successively partitions the data based on the attributes of the study variables, conditioning subsequent division decisions on the possible combinations that the remaining variables can take (Nasteski, 2017; Rokach & Maimon, 2014).

On the other hand, studies on HE dropout using machine learning must address the class imbalance problem, as dropout students represent a smaller proportion of cases than non-dropout students (Palacios et al., 2021), which can lead to errors in the results and model interpretation. To correct this situation, several techniques were employed, including sampling and data boosting, with the RUSboosted procedure, which involves an increase in random undersampling, proving more efficient than other methods based on the Synthetic Minority Oversampling Technique (Chawla et al., 2002; Mduma, 2023).

Finally, for the model with the highest sensitivity and F-score, five thresholds were applied to the variable with the highest entropy, generating alternative classification models. Thresholds one through five consisted of subtracting 0.1 to 0.5 tenths from each student's overall grade point average.

2.6. Testing the models

To determine the effectiveness of each classifier, two procedures were performed. The first consisted of training and calculating the overall accuracy of each classifier using the 10-fold cross-validation method, where the arithmetic mean of the results from each iteration was obtained to determine the overall accuracy. Subsequently, a contingency table with true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) was constructed, and the accuracy, sensitivity, specificity, ROC area, and F-score were determined. For this study, it is fundamental to establish that the sensitivity and F-score measures are the most appropriate for predicting student dropout.

The second procedure involved using the test matrix to evaluate the effectiveness of the predictive model, calculating the same parameters mentioned above.

3. RESULTS

The database for the analysis consisted of 946 students. Dropouts constituted 3.77% of the sample (n=36; 25 females and 11 males). The training matrix had 26 dropout students (3.93% dropout rate), and the test matrix had 10 (3.52% dropout rate). Using the training matrix, the 10-fold cross-validation method was employed to identify the classifier with the best sensitivity to dropout, which was found to be RUSboosted trees. Table 5 presents the classification accuracy, F-score, ROC area, confusion matrix, sensitivity, and specificity for all classifiers.

Table 5. Classification accuracy and confusion matrix.

Classifier	F-score	ACC ^a	Roc Area	TN	FP	TP	FN	Sen	Spec
Complex tree	0.47	95.3%	0.73	623	31	8	18	31%	98%
Medium tree	0.47	95.3%	0.73	623	31	8	18	31%	98%
Simple tree	0.26	95.3%	0.67	627	9	4	22	15%	99%
Linear discriminant	0.67	94.0%	0.88	608	28	14	12	54%	96%
Quadratic discriminant	0.67	93.4%	0.85	604	32	14	12	54%	95%
Logistic Regression	0.37	95.6%	0.85	627	6	6	20	23%	99%
Linear SVM	undefined	95.9%	0.69	635	1	0	26	undefined	>99%
Quadratic SVM	0.42	96.5%	0.78	632	4	7	19	27%	99%
Cubic SVM	0.51	95.3%	0.75	622	14	9	17	35%	98%
Fine Gaussian SVM	undefined	96.1%	0.75	636	0	0	26	undefined	undefined
Medium Gaussian SVM	undefined	95.9%	0.75	635	1	0	26	undefined	>99%

(continued)

Table 5. Classification accuracy and confusion matrix. (*continued*)

Classifier	F-score	ACC ^a	Roc Area	TN	FP	TP	FN	Sen	Spec
Coarse Gaussian SVM	undefined	96.1%	0.75	636	0	0	26	undefined	100%
Fine KNN	0.32	94.4%	0.58	620	16	5	21	19%	97%
Medium KNN	0.08	96.1%	0.76	635	1	1	25	4%	>99%
Coarse KNN	undefined	96.1%	0.81	636	0	0	26	undefined	100%
Cosine KNN	0.37	96.8%	0.84	635	1	6	20	23%	99%
Cubic KNN	0.21	96.4%	0.76	635	1	3	23	12%	>99%
Weighted KNN	undefined	95.9%	0.76	635	1	0	26	undefined	>99%
Boosted trees	0.54	95.9%	0.81	625	11	10	16	38%	98%
Bagged trees	0.42	96.2%	0.87	630	6	7	19	27%	99%
Subspace discriminant	0.51	96.7%	0.86	631	5	9	17	35%	99%
Subspace KNN	0.08	95.9%	0.79	634	2	1	25	4%	>99%
RUSboosted trees	0.83	84.6%	0.90	539	97	21	5	81%	85%

Note: It is important to note that all 23 models were included, which Matlab® allowed us to explore. This tool includes the most popular classification, clustering, and regression algorithms, allowing supervised machine learning to explore data, select characteristics, specify validation schemes, train models, and evaluate results.

^a ACC= Accuracy; FN= False negative; FP= False positive; TN= True negative; TP= True positive; Sen= Sensitivity; Spec= Specificity.

The F-score, as well as the sensitivity, specificity, and accuracy of the RUS-boosted trees classifier, varied according to the number of variables used in the training (Table 6). Following the entropy ranking (Table 4), sets of variables were tested. First, the model was trained considering 20% of the variables (from the first to the second variable, as presented in the entropy ranking, Table 4), then 40% of the variables (from the first to the fourth variable), and so on, until 100% of the variables had been tested. The best accuracy was achieved when using 60% of the variables, while the best combined results for F-score, sensitivity, and specificity were achieved when using 100% of the study variables.

Table 6. Confusion matrix for feature selection using RUSboosted trees.

Features ranked	F-score	ACC ^a	Roc Area	TN	FN	TP	FP	Sen	Spec
1-2 (20% of features)	0.83	84.6%	0.73	539	5	21	97	81%	85%
1-4 (40% of features)	0.83	84.8%	0.87	541	5	21	95	81%	85%
1-6 (60% of features)	0.82	87.1%	0.91	557	6	20	79	77%	88%
1-8 (80% of features)	0.78	84.2%	0.90	539	7	19	97	73%	85%
1-10 (100% of features)	0.84	86.8%	0.92	554	5	21	82	81%	87%

^a ACC= Accuracy; TN= True negative; FN= False negative; TP= True positive; FP= False positive; Sen= Sensitivity; Spec= Specificity.

Using the test matrix with the model based on RUSboosted trees, 80% sensitivity, 75% classification specificity, and an F-score of 0.78 were obtained. Applying thresholds to the first variable of the entropy ranking (grade point average) resulted in 100% sensitivity for detecting dropouts. Model 2 stands out, with an F-score

of 0.80 and a sensitivity of 90%. The performance of the models and the different thresholds used are presented in Table 7.

Table 7. Model testing with new data.

Classifier	F-score	ACC ^a	Roc Area	TN	FN	TP	FP	Sen	Spec
Model 1 RUSboosted trees	0.78	75.4%	0.75	206	2	8	68	80%	75%
Model 2 Umbra1 -0.1	0.80	72.1%	0.71	196	1	9	78	90%	72%
Model 3 Umbra1 -0.2	0.78	68.3%	0.68	185	1	9	89	90%	68%
Model 4 Umbra1 -0.3	0.76	65.1%	0.64	176	1	9	98	90%	64%
Model 5 Umbra1 -0.4	0.74	62.7%	0.62	169	1	9	105	90%	62%
Model 6 Umbra1 -0.5	0.76	61.0%	0.60	163	0	10	111	100%	59%

^a ACC= Accuracy; TN= True negative; FN= False negative; TP= True positive; FP= False positive; Sen= Sensitivity; Spec= Specificity. The thresholds of models 2 to 6 consisted, respectively, of subtracting tenths of the variable ranked with the highest entropy (general grade point average) from 0.1 to 0.5.

4. DISCUSSION

The study posed two research questions and one study aim. The first research question was (i) What is the most effective machine learning classifier for identifying higher education students at risk of dropout using multidisciplinary databases? Of the 23 classifiers trained in this study, RUSboosted Trees presented the best performance in terms of F-score, sensitivity, and specificity, measures that were prioritised over the accuracy of the model, which distinguishes the present work from other studies that have focused their efforts on presenting higher levels of accuracy (Berens et al., 2018; Heredia et al., 2015; Miranda & Guzmán, 2017). The use of the F-score and sensitivity is justified because this measure allows for emphasising the prediction of dropouts, rather than classifying non-dropouts correctly (Barramuño et al., 2022), thereby focusing efforts on identifying students who are at risk of potentially dropping out of university, considering the heterogeneity of the sample. Thus, it can be stated that the RUSBoosted Trees classifier presents the best predictive performance for identifying students at risk of dropout, with results in F-score and sensitivity superior to those of the other 22 classifiers tested.

Consequently, to address the second research question, which was (ii) What is the most effective specific predictive model for identifying higher education students at risk of dropout using multidisciplinary databases?, a process was carried out consisting of applying thresholds to the variable that the literature (Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Berens et al., 2018; Cuji Chacha et al., 2020; Eegdeman, Cornelisz, Meeter et al., 2022; Fahd et al., 2021; Heredia et al., 2015; Kemper et al., 2020; Palacios et al., 2021; Sani et al., 2020; Siri, 2015) and the entropy analysis carried out in this study identified as the one that provides the most significant predictive capacity on the risk of university dropout: the grade point average. It is not surprising that the grade point average holds such significance, as it is an outcome influenced by a multitude of both internal and external factors. While academic aptitude and self-concept (Cancino-Marchant et al., 2022) among other individual variables come into play (Gedda-Muñoz et al., 2023), the pedagogical abilities of educators, the educational environment and university resources, the cultural capital of parents (Bourdieu & Passeron, 2021; Reay, 2022), students' adaptation to the university setting, and various other elements also play pivotal roles. Considering this, within the test matrix, the model employed for classification was based on RUS-boosted trees, incorporating thresholds for the grade point average. Hence, the thresholds on the grade point average were progressively applied to six models based on the RUSBoosted Trees classifier, with model two yielding the best results. The model achieved an accuracy of 72.1%, an F-score of 0.80, a sensitivity of 90%, and a specificity of 72%. The results of the F-score and sensitivity obtained are especially relevant because they demonstrate that this model successfully identifies a greater number of students at risk of dropout. These results are consistent with those reported in other studies, which also indicate similar findings using decision trees to predict the

dropout phenomenon (Kemper et al., 2020; Liang et al., 2016; Márquez-Vera et al., 2016; Ortiz-Lozano et al., 2018; Rokach & Maimon, 2014).

Therefore, this study aimed to develop a multidisciplinary prediction model for dropout risk in higher education students using machine learning. The primary outcomes of the study correspond to a model that allows identifying 90% of dropouts, regardless of their discipline and program of origin, demonstrating and strengthening the literature on the usefulness of machine learning for detecting dropout risk in university students (Barramuño et al., 2022; Berens et al., 2018; Eegdeman, Cornelisz, Meeter et al., 2022; Liang et al., 2016; Márquez-Vera et al., 2016; Mduma, 2023; Sani et al., 2020; Wan Yaacob et al., 2020). Specifically, regarding the performance and predictive capability of the models in the reviewed studies, they present an average accuracy of 85.8%, an F-score of 0.78, a sensitivity of 67.7%, and a specificity of 84.3%. Our study demonstrates improved performance in F-score and sensitivity, which are particularly important for the phenomenon being studied.

Furthermore, it is necessary to point out that several studies using machine learning to investigate university dropouts are based on samples of students from the same programs. This situation affects the transferability of the models and methodologies used in studies conducted with machine learning to other programs or institutions, as their results operate on a homogeneous sample in terms of the disciplinary background of the students. In this sense, the present study presented similar values in terms of precision to those of previous studies (Kemper et al., 2020; Palacios et al., 2021), while also achieving excellent results in terms of sensitivity and F-score. The use of the F-score and sensitivity is justified because this measure allows for emphasising the prediction of dropouts, rather than classifying non-dropouts correctly (Barramuño et al., 2022), thereby focusing efforts on identifying students who are at risk of potentially dropping out of university, considering the heterogeneity of the sample.

This is relevant because the proposed model successfully identifies students who drop out in a multidisciplinary sample, which could be helpful to predict student dropout in HE institutions with characteristics similar to those reported in this study.

Finally, in the field of HE, dropout is presented as a complex phenomenon that involves sociodemographic, individual, and institutional aspects. In this sense, this study included ten variables previously identified as relevant for the analysis and study of university dropout, in line with the theories of Tinto (1975, 2022) and the subsequent studies conducted by Braxton (2019), Donoso and Schiefelbein (2007), Himmel (2002), as well as with the findings of Berens et al. (2018), Cuji Chacha et al. (2020), Fahd et al. (2021), Heredia et al. (2015), Kuz and Morales (2023), and Palacios et al. (2021), in the field of machine learning. These variables were the educational level of parents, age, gender, school origin, years to initiate HE, financing type, number of subjects failed, grade point average, and number of subjects enrolled per year. In terms of the contribution of these variables to predicting educational dropout, it can be highlighted that variables such as the grade point average and the number of subjects failed have substantive importance in the functioning of the machine learning model presented in this study, which is also corroborated by previous research (Abu-Oda & El-Halees, 2015; Barramuño et al., 2022; Berens et al., 2018; Eegdeman, Cornelisz, Meeter et al., 2022; Heredia et al., 2015; Kemper et al., 2020; Palacios et al., 2021; Sani et al., 2020; Siri, 2015) that support the importance of considering academic performance in predicting university dropout. However, this does not impede the consideration of other variables in the study of student dropout, because as Bourdieu and Passeron (2019, 2021) and Reay (2004, 2022) point out in their theories on the reproduction of inequalities in the educational system, there are structural, systemic, and institutional factors that impact the educational trajectories of young people. In this sense, our study included variables such as the educational level of parents and the type of financing used by students (Barramuño et al., 2022; Heredia et al., 2015; Himmel, 2002; Kuz & Morales, 2023; Larroucau, 2016; Palacios et al., 2021; Sani et al., 2020; Siri, 2015), and although these elements did not appear as especially significant for predicting dropout using machine learning models, they must be persistently incorporated into analyses of this phenomenon, such as gender, high school background or others, because the inclusion of such variables can enhance our understanding of social inequalities in the university context.

Among the study's limitations: (i) the voluntary participation of students restricted the amplitude of the sample and involved a larger number of female students; (ii) due to the data collection method employed, variables related to students' previous academic performance (i.e., upper high school grades) were not included, which could have provided valuable insights into their likelihood of dropout; (iii) socioeconomic status defining variables such as parental occupation and monthly family income were not incorporated, which could have shed light on the socioeconomic factors influencing dropout rates. However, the 'gratuity' and the parents' educational level variables utilised served as an indirect measure of this.

Some of the strengths of this study include: (i) the number of the sample, consisting of 946 students; (ii) the diversity of knowledge areas represented in the sample, which provides practical utility, since other HE institutions could replicate it; (iii) It is noted that currently, machine learning enables the development of highly

accurate models for predicting university student dropout. However, challenges remain in transferring these models across different institutions and programs. While the necessary technology and procedures exist, widespread adoption by institutions has yet to occur. This limitation may arise because most predictive models are based on samples from specific fields of study or academic degrees, thereby limiting their applicability beyond those contexts. In contrast, the model developed in this study draws from students across various disciplines, establishing a multidisciplinary database. This approach enables its implementation across entire institutions, demonstrating its efficacy in accurately predicting student dropout rates. This study aims to promote the implementation of a predictive model that is useful in various contexts and easily transferable between institutions, thereby encouraging the adoption of this technological tool.

In this sense, the work developed is auspicious for at least four reasons: (i) the proposed model has successfully identified dropouts, operating on a multidisciplinary student sample; (ii) the variables used in this research are generally available in academic records of HE institutions, therefore, the model we propose represents a tool that universities can use to implement or complement student dropout prevention programs, leveraging their own academic record data; (iii) the substantiated presentation of the F-score and sensitivity as the ideal metrics to interpret the effectiveness of a predictive model of student dropout, over other metrics such as accuracy or precision; (iv) this study has opened up possibilities to further the use of machine learning in educational environments, specifically, e.g., to address the university dropout phenomenon, introducing new attributes to the work developed so far, such as the socioeconomic origin of students, the vocational aspects associated with the process of enrolling in study programs, and the identification of differences between entrance cohorts.

5. CONCLUSION

Based on the comprehensive analysis conducted in this study, several findings have emerged. Firstly, the utilisation of the RUSboosted Trees classifier has demonstrated superior predictive performance in identifying students at risk of dropout, surpassing 22 other classifiers examined. By prioritising measures such as F-score and sensitivity over accuracy, this approach makes a relevant contribution to conventional studies, which often focus solely on achieving higher accuracy rates. The proposed predictive model with the highest F-score and sensitivity corresponds to the RUS-boosted trees classifier, with a threshold of -0.1 in the grade point average, successfully recognising 90% of the dropouts, even under the multidisciplinary nature of the programs from which the students originated. These outcomes align with existing literature on the significance of academic performance indicators in predicting student attrition, while also emphasising the importance of considering diverse disciplinary backgrounds within predictive modelling frameworks. The institutions will be able to use the variables identified in this study and the machine learning procedures described to improve their institutional policies aimed at promoting educational success and reducing university dropouts.

ACKNOWLEDGMENT

The authors are grateful for the students' participation in this study.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

FUNDING

The author received no financial support for the research or authorship of this article.

AVAILABILITY OF MATERIALS

Anonymised data used for this research are available on request from the corresponding author, who will consider each request. The script that enables visualisation, calculations, and procedures developed is available in the following repository.

AUTHOR'S CONTRIBUTIONS

Mauricio Barramuño-Medina: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing.

Pablo Valdés-Badilla: Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – original draft, Writing – review.

José Luis Saavedra-Lucero: Conceptualization, Investigation, Resources, Writing – original draft.

Germán Gálvez-García: Methodology, Supervision, Validation, Writing – review.

Relmu Gedda-Muñoz: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Project administration, Writing – original draft, Writing – review and editing.

REFERENCES

- Abu-Oda, G., & M. El-Halees, A. (2015). Data Mining in Higher Education: University Student Dropout Case Study. *International Journal of Data Mining & Knowledge Management Process*, 5(1), 15–27. <https://doi.org/10.5121/ijdkp.2015.5102>
- Aljohani, O. (2016). A Review of the Contemporary International Literature on Student Retention in Higher Education. *International Journal of Education and Literacy Studies*, 4(1), 40–52. <https://doi.org/10.7575/aiac.ijels.v4n.1p.40>
- Aulck, L., & West, J. (2017). Attrition and performance of community college transfers. *PloS One*, 12(4), e0174683. <https://doi.org/10.1371/journal.pone.0174683>
- Barramuño, M., Meza-Narváez, C., & Gálvez-García, G. (2022). Prediction of student attrition risk using machine learning. *Journal of Applied Research in Higher Education*, 14(3), 974–986. <https://doi.org/10.1108/JARHE-02-2021-0073>
- Beer, C., & Lawson, C. (2016). The problem of student attrition in higher education: An alternative perspective. *Journal of Further and Higher Education*, 41(6), 773–784. <https://doi.org/10.1080/0309877x.2016.1177171>
- Berens, J., Oster, S., Schneider, K., & Burghoff, J. (2018). Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data and Machine Learning Methods. *Schumpeter Discussion Papers*, 6(1), 1–38. <https://doi.org/10.2139/ssrn.3275433>
- Bourdieu, P., & Passeron, J. C. (2019). *La reproducción: Elementos para una teoría del sistema educativo*. Siglo XXI Editores.
- Bourdieu, P., & Passeron, J. C. (2021). *Los Herederos: los estudiantes y la cultura*. Siglo XXI & Clave Intelectual.
- Braxton, J. M. (2019). Leaving College: Rethinking the Causes and Cures of Student Attrition by Vincent Tinto. *Journal of College Student Development*, 60(1), 129–134. <https://doi.org/10.1353/csd.2019.0012>
- Cancino-Marchant, A., Gedda-Muñoz, R., Herrera-Valenzuela, T., Magnani Branco, B., Guzmán-Muñoz, E., López-Fuenzalida, A., & Valdés-Badilla, P. (2022). Autoestima y autoconcepto académico en escolares durante la pandemia por la COVID-19. *Revista Cubana de Medicina Militar*, 51(3), e02202012.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Choudhary, N., Shinde, Y., Kannan, R., & Venkatraman, V. (2014). Impact of attribute selection on the accuracy of Multilayer Perceptron. *International Journal of IT & Knowledge Management*, 7(2), 32–36.
- Christou, V., Tsoulos, I., Loupas, V., Tzallas, A. T., Gogos, C., Karvelis, P. S., Antoniadis, N., Glavas, E., & Giannakeas, N. (2023). Performance and early drop prediction for higher education students using machine learning. *Expert Systems with Applications*, 225, 120079. <https://doi.org/10.1016/j.eswa.2023.120079>
- CNED. (2022). *Base de datos índices CNED 2005-2022*. Consejo Nacional de Educación, Ministerio de Educación-Chile.
- Cuji Chacha, B. R., Gavilanes López, W. L., Vicente Guerrero, V. X., & Villacis Villacis, W. G. (2020). Student Dropout Model Based on Logistic Regression. In *Communications in Computer and Information Science* (pp. 321–333). https://doi.org/10.1007/978-3-030-42520-3_26
- Donoso, S., & Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social. *Estudios Pedagógicos (Valdivia)*, 33(1). <https://doi.org/10.4067/s0718-07052007000100001>

- Dutt, A. (2015). Clustering Algorithms Applied in Educational Data Mining. *International Journal of Information and Electronics Engineering*, 5(2), 112–116. <https://doi.org/10.7763/ijee.2015.v5.513>
- Eegdeman, I., Cornelisz, I., Meeter, M., & van Klaveren, C. (2022). Identifying false positives when targeting students at risk of dropping out. *Education Economics*, 31(3) 313–325. <https://doi.org/10.1080/09645292.2022.2067131>
- Eegdeman, I., Cornelisz, I., van Klaveren, C., & Meeter, M. (2022). Computer or teacher: Who predicts dropout best? *Frontiers in Education*, 7, 1–10. <https://doi.org/10.3389/FEDUC.2022.976922/BIBTEX>
- Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2021). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, 2021, 1–33. <https://doi.org/10.1007/S10639-021-10741-7>
- Frank, E., Hall, M., & Witten, I. (2017). The WEKA workbench. In I. Witten, E. Frank, M. Hall, & C. Pal (Eds.), *Data Mining: Practical Machine Learning Tools and Techniques* (pp. 553–571). Elsevier, Morgan Kaufmann. <https://doi.org/10.1016/B978-0-12-804291-5.00024-6>
- Garbanzo Vargas, G. M. (2012). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, 31(1), 43. <https://doi.org/10.15517/revedu.v31i1.1252>
- Gedda-Muñoz, R., & Carrasco-Bahamonde, J. A. (2023). Inclusión y acceso en universidades chilenas: El caso de la Política de Gratuidad. *Revista Educación Superior y Sociedad (ESS)*, 35(1), 365–395. <https://doi.org/10.54674/ESS.V35I1.630>
- Gedda-Muñoz, R., Fuentez Campos, Á., Valenzuela Sakuda, A., Retamal Torres, I., Cruz Fuentes, M., Badicu, G., Herrera-Valenzuela, T., & Valdés-Badilla, P. (2023). Factors Associated with Anxiety, Depression, and Stress Levels in High School Students. *European Journal of Investigation in Health, Psychology and Education*, 13(9), 1776–1786. <https://doi.org/10.3390/ejihpe13090129>
- Gedda-Muñoz, R., & Villagrán-Del Picó, N. (2020). Academic self-concept in university students: Their association with parents' educational level and previous experience in higher education. *Journal of Educational, Cultural and Psychological Studies*, 2020(22), 109–124. <https://doi.org/10.7358/ecps-2020-022-gedd>
- Gnanambal, S., Thangaraj, M., Meenatchi V, & Gayathri, V. (2018). Classification Algorithms with Attribute Selection: an evaluation study using WEKA. *International Journal of Advanced Networking and Applications*, 9(6), 3640–3644.
- Harrington, P. (2012). *Machine Learning in Action*. Manning Publications.
- Heredia, D., Amaya, Y., & Barrientos, E. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *IEEE Latin America Transactions*, 13(9), 3127–3134. <https://doi.org/10.1109/tla.2015.7350068>
- Himmel, E. (2002). Modelo de análisis de la deserción estudiantil en la educación superior. *Calidad en la Educación*, 17, 91. <https://doi.org/10.31619/caledu.n17.409>
- Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, 10(1), 28–47. <https://doi.org/10.1080/21568235.2020.1718520>
- Kuz, A., & Morales, R. (2023). Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México. *Education in the Knowledge Society*, 24, e30080. <https://doi.org/10.14201/eks.30080>
- Larroucau, T. (2016). Estudio de los factores determinantes de la deserción en el sistema universitario chileno. *Revista Estudios de Políticas Públicas*, 1. <https://doi.org/10.5354/0719-6296.2015.38351>
- Liang, J., Yang, J., Wu, Y., Li, C., & Zheng, L. (2016). Big Data Application in Education: Dropout Prediction in Edx MOOCs. *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, 440–443. <https://doi.org/10.1109/BigMM.2016.70>
- Limsathitwong, K., Tiwatthanont, K., & Yatsungnoen, T. (2018). Dropout prediction system to reduce discontinue study rate of information technology students. *2018 5th International Conference on Business and Industrial Research (ICBIR)*, 110–114. <https://doi.org/10.1109/ICBIR.2018.8391176>
- Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Fardoun, H. M., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1), 107–124. <https://doi.org/10.1111/exsy.12135>
- Mduma, N. (2023). Data Balancing Techniques for Predicting Student Dropout Using Machine Learning. *Data*, 8(3), 49. <https://doi.org/10.3390/data8030049>
- Miranda, M. A., & Guzmán, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. *Formación Universitaria*, 10(3), 61–68. <https://doi.org/10.4067/S0718-50062017000300007>
- Nasteski, V. (2017). An overview of the supervised machine learning methods. *HORIZONS.B*, 4, 51–62. <https://doi.org/10.20544/HORIZONS.B.04.1.17.P05>

- OECD. (2020). *Education at a Glance 2020: OECD Indicators*. <https://doi.org/10.1787/69096873-en>
- OECD. (2023). *Education at a Glance 2023*. OECD. <https://doi.org/10.1787/e13bef63-en>
- OFD-CIAE. (2021). *Informe N°9 Observatorio Formación Docente: Deserción en carreras de Pedagogía en Chile* (9).
- Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., & Casadesús-Fa, M. (2018). University student retention: Best time and data to identify undergraduate students at risk of dropout. *Innovations in Education and Teaching International*, 57(1), 74–85. <https://doi.org/10.1080/14703297.2018.1502090>
- Palacios, C. A., Reyes-Suárez, J. A., Bearzotti, L. A., Leiva, V., & Marchant, C. (2021). Knowledge Discovery for Higher Education Student Retention Based on Data Mining: Machine Learning Algorithms and Case Study in Chile. *Entropy*, 23(4), 485. <https://doi.org/10.3390/e23040485>
- Parra-Sánchez, J. S., Torres Pardo, I. D., & Martínez de Merino, C. Y. (2023). Factores explicativos de la deserción universitaria abordados mediante inteligencia artificial. *Revista Electrónica de Investigación Educativa*, 25, 1–17. <https://doi.org/10.24320/REDIE.2023.25.E18.4455>
- Pitman, T., Koshy, P., & Phillimore, J. (2014). Does accelerating access to higher education lower its quality? The Australian experience. *Higher Education Research & Development*, 34(3), 609–623. <https://doi.org/10.1080/007294360.2014.973385>
- Reay, D. (2004). 'It's all becoming a habitus': beyond the habitual use of habitus in educational research. *British Journal of Sociology of Education*, 25(4), 431–444. <https://doi.org/10.1080/0142569042000236934>
- Reay, D. (2022). 'The more things change the more they stay the same': The continuing relevance of Bourdieu and Passeron's Reproduction in Education, Society and Culture. *Revista Española de Sociología*, 31(3), a116. <https://doi.org/10.22325/FES/RES.2022.116>
- Rokach, L., & Maimon, O. (2014). *Data mining with decision trees: Theory and applications*. World Scientific.
- Sani, N. S., Fikri, A., Ali, Z., Zakree, M., & Nadiyah, K. (2020). Drop-Out Prediction in Higher Education Among B40 Students. *International Journal of Advanced Computer Science and Applications*, 11(11). <https://doi.org/10.14569/IJACSA.2020.0111169>
- Santelices, M., Horn, C., Catalán, X., & Venegas, A. (2021). Aggregated results of access programs implemented by universities in Chile: Students' persistence using a matched sample. *Higher Education Policy*, 35, 498–521. <https://doi.org/10.1057/s41307-021-00223-3>
- SIES. (2019). *Deserción de primer año y Reingreso a la Educación Superior en Chile*. <https://d66z.short.gy/K7cjsq>
- SIES. (2023). *Informe Retención SIES 2023*. Servicio de Información de Educación Superior, Ministerio de Educación. <https://d66z.short.gy/7uEvS3>
- Siri, A. (2015). Predicting Students' Dropout at University Using Artificial Neural Networks. *Italian Journal of Sociology of Education*, 7(2), 225–247.
- Soto-Hernández, V. J. (2016). Estudiantes de primera generación en Chile: una aproximación cualitativa a la experiencia universitaria. *Revista Complutense de Educación*, 27(3). https://doi.org/10.5209/rev_RCED.2016.v27.n3.47562
- Teixeira, P., Kim, S., Landoni, P., & Gilani, Z. (2017). *Rethinking the public-private mix in Higher Education. Global trends and national policy challenges*. Brill. <https://doi.org/10.1007/978-94-6300-911-9>
- Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, 45(1), 89–125. <https://doi.org/10.3102/00346543045001089>
- Tinto, V. (2022). Increasing Student Persistence: Wanting and Doing. In H. Huijser, M. Y. Kek, & F. Padró (Eds.), *Student Support Services. University Development and Administration* (pp. 53–70). Springer. https://doi.org/10.1007/978-981-16-5852-5_33
- Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers & Education*, 143, 103676. <https://doi.org/10.1016/j.compedu.2019.103676>
- Wan Yaacob, W. F., Mohd Sobri, N., Nasir, S. A. M., Norshahidi, N. D., & Wan Husin, W. Z. (2020). Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques. *Journal of Physics: Conference Series*, 1496, 012005. <https://doi.org/10.1088/1742-6596/1496/1/012005>
- Zwick, T. (2012). Determinants of Individual Academic Achievement - Group Selectivity Effects Have Many Dimensions. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2185386>