Revista Educación, 2025, 49(1), enero-junio

ISSN: 0379-7082 / e-ISSN 2215-2644

Scientific research article

DOI: http://doi.org/10.15517/revedu.v49i1.61275

Student dropout from a General Mathematics Course: Identification of Relevant Predictive Variables

Abandono estudiantil en el curso de Matemática General: identificación de variables relevantes para su predicción

José Andrey Zamora-Araya
Universidad Nacional
Heredia, Costa Rica
jzamo@una.ac.cr (Correspondencia)
https://orcid.org/0000-0001-6050-5850

Tania Elena Moreira-Mora
Instituto Tecnológico de Costa Rica
Cartago, Costa Rica
tmoreira.costarica@gmail.com
https://orcid.org/0000-0002-8955-0804

English Translation:
Xinia Rodríguez Castillo

Received: 2 September 2024
Accepted: 30 October 2024

¿How to cite this article?

Zamora-Araya, J. A. y Moreira-Mora, T. E. (2025). Student dropout from a General Mathematics Course: Identification of Relevant Predictive Variables. *Revista Educación, 49*(1). http://doi.org/10.15517/revedu.v49i1.61275

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

ABSTRACT

The aim of this study was to determine the most important variables for predicting student dropout from the General Mathematics course (MAT001) of the Universidad Nacional de Costa Rica (UNA), considering the types of students and the time at which dropout takes place. Six predictive models were constructed (two student groups at three different times) and three supervised learning algorithms were implemented in each model: Logistic Regression (LR), Random Forest (RF) and XGBoost (XGB). The total sample was split into training files containing data on students who enrolled in the course during the years 2017 and 2018, and test files with data corresponding to students who enrolled in the year 2019. Once the hyperparameters were fitted (10-fold validation), the main variables associated with student dropout (SD) in the General Mathematics course of each model were identified based on the Gini importance measure; performance of the algorithms ranged from F1-Scores of 0.6251 to 0.7300. In addition, the predictive power of the algorithms in each model were compared by means of a repeated-measures ANOVA with 10-fold cross-validation, and no significant differences were found between the three algorithms in any of the proposed models. The main variables associated with student dropout (SD) were academic, such as grades on the academic attitude test (AAT), high school education grades, and grades on MAT001 tests, student attributes as sex and age at enrollment, economic factors such as scholarships and the Social Development Index (SDI), and institutional factors such as high school educational opportunities that students were exposed to, and the ages and specializations of the teaching staff. Based on the results of this analysis, it is recommended that teachers specialized in Educational Mathematics be assigned to teach initial courses, and to propose designs for decision making about actions that increase permanence.

KEYWORDS: Higher Education, Student Dropout, Student Performance, Statistical Data, Statistical Methodology, Mathematics.

RESUMEN

El objetivo del estudio fue determinar las variables más importantes para la predicción del abandono estudiantil del curso de Matemática General, de la Universidad Nacional de Costa Rica (UNA), considerando el tipo de estudiante y el momento en que se identifica el abandono. Se construyeron seis modelos predictivos (dos grupos estudiantiles en tres momentos diferentes) y en cada modelo se implementaron tres algoritmos de aprendizaje supervisado: Regresión Logística (RL), Random Forest (RF) y XGBoost (XGB). La muestra total se dividió en archivos de entrenamientos, estudiantado que matriculó el curso durante los años 2017 y 2018 y archivos de prueba correspondientes a la matrícula del año 2019. Una vez calibrados los hiperparámetros (validación 10 folds) se identificaron las principales variables asociadas con abandono estudiantil (AE) en el curso de Matemática General de cada modelo con base en

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

la medida de importancia de Gini. No obstante, el rendimiento de los algoritmos oscila entre valores de F1- Score de 0.6251 y 0.7300. Además, se comparó el poder predictivo de los algoritmos en cada modelo por medio de un ANOVA de medidas repetidas con validación cruzada con 10 Folds, y no se encontraron diferencias significativas entre los tres algoritmos en ninguno de los modelos propuestos. Las principales variables asociadas al abandono estudiantil (AE) son de tipo académico como la nota de la prueba de actitud académica (PAA), la nota de colegio y las notas de las pruebas parciales, individuales; como el sexo y la edad de ingreso, económicas; como la beca y el índice de desarrollo social (IDS) e institucionales como el estrato, la edad y especialización del personal docente. Se recomienda asignar al profesorado especializado en Matemática Educativa para impartir los cursos iniciales y el diseño propuesto para tomar decisiones sobre acciones que aumenten la permanencia.

PALABRAS CLAVE: Enseñanza superior, Abandono escolar, Rendimiento escolar, Datos estadísticos, Metodología estadística, Matemáticas.

INTRODUCTION

The phenomenon of student dropout (SD) has been studied since 1940, initially focusing on desertion. From the 1990s onwards, with the massification of tertiary education, the problem became more apparent, since the student body entering university classrooms was more heterogeneous and was under pressure to obtain a university degree or a certification that would provide greater possibilities of participation in a competitive and demanding labor market (Wang et al., 2023).

University SD can be analyzed at the level of the educational system, the institution, or the curricula in which the students enroll, the latter two being those in which most studies on SD are focused. More recently, interest has been raised in researching SD in particular subjects or courses, among which those with substantial mathematical content often show a particularly high level of SD, which usually occurs during the first year of university study. For example, in the case of the General Mathematics course (MAT001) of the Universidad Nacional (UNA), excluding the pandemic years, on average less than a third of students manage to complete the course (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020).

Among the repercussions of SD at the course level are increased duration of the courses and increased costs caused by the need to offer courses for students repeating the course. In addition, it makes it difficult for the administration to meet student demand for places in this course in a scenario of limited budgetary resources.

According to the State of the Nation Program (PEN), dropout rates in Costa Rica for public higher education vary between 43.4% and 65.3% (Román, 2017). Higher education institutions are therefore deeply concerned with proposing strategies that increase retention levels, and strive to expand the coverage and quality of the academic programs offered.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Regarding the use of data analysis techniques, research on the topic in Costa Rica has focused on the analysis of factors associated with SD in descriptive, correlational studies, generally associated with entry cohorts or specific programs of study (Pascua-Cantero, 2016; Rodríguez-Pineda y Zamora-Araya, 2014), in which it has been found that age of entry, sex, scholarship and academic variables were relevant in explaining SD.

Recently, more sophisticated statistical techniques such as survival analysis (Mora, 2016) and the application of predictive models (Solís et al., 2018), have been applied in Costa Rica, but there are few studies that analyze the phenomenon from the perspective of predicting SD at the level of subjects or courses in the area of mathematics (Zamora-Araya, 2023a), since most are carried out at the level of degree programs or student cohorts.

Internationally, studies have been carried out that focus on Mathematics courses, but usually with a more explanatory than predictive approach. For instance, Muñoz-Camacho et al. (2018) carried out a logistic regression analysis to determine the probabilities of SD in students enrolled in Basic Mathematics courses in the period from 2011 to 2016. The results showed correct classification rates of between 59% and 70% depending on the university department or school students were enrolled in. They also found that being enrolled in a mathematics course offered during a regular academic semester instead of an intensive course (in summer school, for instance), being enrolled in online classes with few formal requirements, being in classes of more than 50 students, being a male, and being enrolled in the program of studies of the Faculty of Economic Sciences all increased the probability of SD.

Calva et al., (2021) in another study of academic success in classroom Mathematics courses, measured in terms of passing the course, conducted an analysis to predict passing in remedial courses (Mathematics, Physics, Chemistry and Language). It was found that the use of gradient boosting techniques in their analysis produced the best results, and that the variables with the greatest influence were the grade obtained on the first course test, the weighted grade average, the student's university admission test grade, the major area of study chosen by the student, and the number of courses students were enrolled in; an accuracy of 89.1% and an area under the curve (AUC) of 95.5% were obtained in this study. Other relevant variables suggested by the Logistic Regression model were student age, place of residence, marital status, number of family members, type of school, time of enrollment, student population segment (affirmative action admission, territorial merit admission, general population), semester (first or second) and shift (daytime or evening classes).

In an investigation of passing courses by students in the MINT disciplines (pure mathematics, information science, natural science and technology), the results of students in these courses were compared with those of students who chose a major in Mathematics teaching and took the course in the 2014-2015 academic year (Kilian et al., 2020). Techniques used included binary logistic regression algori-

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

thms, logistic regression with elastic-net regularization, and Support Vector Machine and tree-based methods, using metrics such as Cohen's Kappa and f1 to evaluate accuracy, precision, and recall.

The results showed values of the F1 metric between 0.60 and 0.73 with a correct prediction of outcomes in the course for 75% of the students. In addition, the most relevant predictors were previous performance grades in Mathematics courses, general school performance, whether students were in their first semester of Mathematics or later semesters, and the modality (classroom or online) of secondary education. Neither Kilian et al. (2020) nor Calva et al. (2021) analyzed SD as a dependent variable, but rather as variables associated with academic performance such as passing and final grades in the course.

International and national studies recognize the importance of using tools such as predictive models for the early detection of students with a high probability of dropping out. The objective of this investigation is to determine relevant variables for correctly predicting the probability of SD of students enrolled in the General Mathematics course at the Universidad Nacional de Costa Rica (UNA), considering the type of student and the moment in which SD occurs.

Most of the research related to SD in Mathematics courses is limited to environments with some degree of virtuality (Kilian et al., 2020; López-Zambrano et al., 2021) due to the ease of collecting data through computer platforms. Therefore, one of the major contributions of this study is the use of supervised learning techniques to analyze SD in an introductory level Mathematics course in a traditional classroom modality, with relatively large samples of students, including students who enrolled in the course in their first semester in the university as well as students who enrolled in the course in later semesters. This distinction between early and later enrollment in a mathematics class has not been made in other investigations of this subject.

Empirical evidence has also shown that academic, economic and personal variables have been associated with SD in the General Mathematics course (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020). However, other variables related to attributes of the teaching staff, educational environments, characteristics and evaluations of the courses, together with traditional academic variables such as credits earned, weighted grade averages, and diagnostic evaluations have also been used to implement predictive models in the educational fieldo (Xu et al., 2022). Hence the importance of incorporating institutional variables associated with SD in this study.

This document is organized into the following sections: firstly, a justification of the problem, the national and international backgrounds of SD, conceptual references of the variables associated with SD, and a description of the use of data mining in the educational field. The methodological section then describes the research design, the population, exclusion criteria and the process of separating the files for training and validation, as well as the design of the proposed models and the analysis strategy. A discussion of results is then presented, followed by the section of conclusions and recommendations.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Conceptual References

Definitions of student dropout (SD)

Studies of student dropout (SD) have addressed conceptual aspects of the term (Cabrera et al., 2006; Tinto, 1989), tipos de abandono (Behr et al., 2020; Munizaga et al., 2018), associated factors, theoretical models to predict or explain the phenomenon (Khoushehgir y Sulaimany, 2023; Wang et al., 2023) and the consequences it generates, which shows the multidimensional and complex nature of both the phenomenon and the term.

Proof of this complexity can be seen in the fact that the term is often equated with others such as school desertion, absenteeism, school failure or educational exclusion, although they do not all represent the same construct, nor do all people use the terminology in the same way to refer to the phenomenon (Zamora-Araya et al., 2023). This reflects the lack of agreement in the scientific community on the subject, and researchers often adopt their own definition that responds to their own needs and interests (Valencia et al., 2024). As Tinto (1982) points out, "it is likely that no definition of dropout fully captures the complexity of its appearance in higher education" (p. 14).

The theoretical definition closest to the approach used in the present investigation is that of Bäulke et al. (2022), who view SD as a decision-making process, since the final decision of students to leave the university can be affected, in part, by conditions at the time of entering the university, including thoughts and intentions about SD. On the other hand, given the absence in the literature of an operational definition of the SD associated with a specific course or subject presented in a classroom modality, the approach taken in this study has been influenced by what Educational Data Mining (EDM) studies have suggested for the early detection of SD, specifically, the use of variables from e-learning system records, student information systems, surveys, course evaluations and online access records (Xu et al., 2022). Thus, it is assumed in this study that a person drops out of the MAT001 course if they do not take the last of the three tests in this course. Adapting the classification proposed by Castaño et al. (2004) and Munizaga et al. (2018), it is said that dropout is very early if a student does not take any of the three course tests, early if they take only the first test, and late if they only take the first two tests.

Variables associated with SD

SD can be analyzed using different approaches, including sociological, interactionist, organizational, psychological, economic and integrative (Lázaro et al., 2020). The variables that are most commonly used in these approaches can be categorized into four areas: individual, socioeconomic, academic and institutional (Guzmán et al., 2021; Zamora-Araya, 2023a). Therefore, this investigation is structured around a categorization of variables rather than of factors.

In accordance with the multidimensionality and diverse causes of SD, the Alfa project for Comprehensive University Management of Dropout (Alfa-GUIA), currently the GUIA Network, proposed the following model with its respective variables to study dropout (Proyecto ALFA-GUÍA, 2013).

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

• Individual: refers to personal characteristics of the student, such as the loss of motivation that influences dropout, which is in turn associated with the topic of achievement behaviors. Associated variables include vocation, economic dependence, motivation, study habits, and adaptation to university life.

• Academic: including the students' institutions of origin, scores on tests prior to entering the university, prior knowledge, study habits, number of credits enrolled for, and other variables that affect performance and in turn student dropout.

• Economic: incorporates variables related to monetary aspects and use of economic resources such as family and personal income, ability to finance studies, and satisfaction upon completing a degree.

• Cultural: linked to the beliefs, values and practices that are part of the student's cultural context, which can affect their emotional stability and motivation.

• Institutional: associated with variables related to scholarships, quality of teaching, and psychosocial care, together with characteristics and integration of the teaching staff with the student body, among others.

SD and Data Mining in Education

During the last decade, SD studies have incorporated predictive models as a tool for the analysis of student academic success. This trend of applying data mining and Machine Learning (ML) techniques to problems related to learning activities in educational environments is part of so-called educational data mining (EDM).

Among the most important variables used in predicting academic performance in EDM are cumulative grade point average, internal assessments, and, to a lesser extent, the students' sociodemographic variables (sex, socioeconomic status, poverty level, age, scholarship, parents' occupation), extracurricular activities (sports, artistic, cultural), high school background (grades, admission tests), and the social interaction network ,such as cell phone use and activity on platforms and social networks (Albreiki et al., 2021; Khan y Ghosh, 2021; Khoushehgir y Sulaimany, 2023).

Machine Learning techniques have been applied in online and classroom courses, where the dependent variable is usually the final grade for the subject or the level of students passing (or failing) the course, instead of SD (Alvarado y Zambrano, 2020; Calva et al., 2021).

In the case of courses in the area of Science and Mathematics, studies use information on academic subjects either as independent variables that help predict graduation (Opazo et al., 2021) or as dependent variables, seeking to predict grades, passing level and, to a lesser extent, permanence or SD in the course (Alvarado y Zambrano, 2020; Kilian et al., 2020). Shin y Shim (2021) point out that studies pay more attention to identifying important factors that influence student performance in Mathematics or Science than to predicting student performance; the relatively few studies that do attempt to predict performance collect data through platforms in online courses to obtain information on what students do or do not do.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

METHODOLOGY

Type of Research

A correlational predictive cohort study is proposed, focusing on the study of events that will occur in the future (Hernández et al., 2014). Three supervised learning algorithms were used to make predictions: Logistic Regression (LR), Random Forest (RF) and XGBoost (XGB).

Population

Data for three cohorts of students who enrolled in the MAT001 course at UNA between 2017 and 2019 were considered – that is, the entire student population who, during the years 2017, 2018 and 2019, enrolled in the MAT001 course at least once, either in the first or second academic semester of one of these years. The original data file contained 5906 records, but after applying exclusion criteria, 91 cases were rejected, leaving a total of 5815 records available for analysis.

Exclusion criteria

Exclusion of student records was based on the following criteria: (a) the student withdrew for justified reasons, (b) the student could not be located or had no identification, (c) there was no record of test grades or incomplete records, (d) the student took only the tests, to demonstrate proficiency in the subject matter, and (e) there was incomplete information about the admission process, such as missing data about student admission grades, students' high school stratum[1] and high school grades. The use of Criterion (e), above, excluded students who entered the UNA under special agreements or modalities, as well as those who entered before 2009, since the admissions system only began to stratify applicants for admission to the UNA in that year.

Collection of information

Data was provided by two sources: the coordinator of Service Courses of the School of Mathematics, which provided records of the teaching staff and grades for the course for 2017, 2018 and 2019, and the university's Registration Department, which provided information related to the admission process, academic performance and socioeconomic variables in an official document, which stipulated the requirements for confidentiality and management of information in accordance with the criteria of the document UNA-AJ-DICT-17-2020, issued by the UNA Legal Advisory Department, in accordance with Law 8968 on the protection of personal data (Asamblea Legislativa, 2011).

Separating training and validation files

In Machine Learning models, the data used is usually divided into two sets: one for training and one for testing. It is common to randomly take a percentage of the total data available for training (from 70% to 90%) and use the rest as test data.

[1] Costa Rican high schools can be divided into three categories, or strata. Stratum 1 consists of private, scientific and experimental bilingual schools, Stratum 2 consists of public academic and technical schools, and Stratum 3 consists of other educational modalities such as night schools, rural schools, distance education, etc. (Aguilar-Fernández et al., 2024)

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

In this investigation, the training file contained the data for students who enrolled in the course during 2017 and 2018, while the 2019 enrollment data was set aside as a test file. This means that neither the training data set nor the validation data set was randomly selected, since the year the course was taken was used as the separation criterion.

This method of dividing the training and test sets is a strength, since it represents a realistic situation in which predictive models would be used – predicting the SD of an entering cohort of MAT001 students based on information available from previous academic cycles, university records, and data gathered as a semester progresses, as has been done in other studies (Kilian et al., 2020).

Description of design

Data for the student population was divided into two groups according to admission status. The first group is made up of individuals from the 2017, 2018, and 2019 cohorts who enrolled in the MAT001 course during the first semester of their first year at the university. For this group, information is only available for variables associated with the admission process and the course itself, such as grades from MAT001 tests and institutional information related to the characteristics of the course's teaching staff. The second group consists of students who enrolled in the MAT001 course between the first semester of 2017 and the second semester of 2019, but who had at least one previous semester of classes at the university before enrolling in the MAT001 course, including students who entered in years prior to 2017. In addition to information concerning the admission process and the course, data on members of this second group includes a university academic history that increases the number of variables available for prediction, such as weighted grade average or the number of subjects dropped in the previous semester.

The time-based classification of dropout proposed by Castaño et al. (2004) was considered when designing the analysis:

1. Very early: includes students who, having completed all the procedures for admission, do not complete the process of enrollment.

2. Early: includes students who drop out of school during the first four semesters of the degree program (depending on the course, the number of semesters may vary).

3. Late: includes students who drop out of school after the early dropout period.

Adopting this approach in the present investigation, the following categories of SD were defined using the number of tests taken in the MAT001 course as an indication of the temporal dimension of SD: (1) very early for students who enrolled in the course, but did not take any of the three tests in the course, and in most cases did not even attend classes; (2) early for those who took the first test, but did not take the second or third test; and (3) late for those who took the first two tests, but not the third.

By jointly considering both populations (those who enrolled in the course in their first semester at the university, and those who enrolled in a later semester) and the time categories (very early dropout,

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

early dropout, and late dropout), six models were designed to analyze SD: (1) first semester enrollees - no tests taken, (2) first semester enrollees - only first test taken, (3) first semester enrollees - first two tests taken, (4) later semester enrollees - no tests taken, (5) later semester enrollees - only first test taken, (6) later semester enrollees – only first two tests taken.

The application of this design constitutes a methodological contribution to studies on SD, since research in this area generally does not consider these distinctions when making predictions, or considers other characteristics associated with the area of knowledge of the career, sex, or institutional affiliation (Albreiki et al., 2021; Khan y Ghosh; 2021; Kilian et al., 2020).

In addition, three different supervised learning algorithms were run on the six models: (1) Logistic Regression (LR), (2) Random Forest (RF), and (3) XGBoost (XGB). The reason for using these algorithms is to compare the results of LR, considered a white-box algorithm due to its advantage of knowing in detail how results are obtained and the ease of interpreting these results, with the other two black-box algorithms, one a bagging algorithm and the other a boosting algorithm, which, despite their greater complexity both in terms of interpretation and processing, have the potential to provide better predictive performance. Also, both to compare the performance of the algorithms and to adjust their hyperparameters, the 10-fold cross-validation method was used, which internally performs a random selection of instances to perform the validation in each fold.

Table 1 shows the number of students in each training and test file per model.

Table 1.

Number of records and percentage of SD for each predictive model of the MAT001 course according to the number of tests taken and semester of enrollment for the 2017-2019 training and test files

Number of tests taken	First semester enrollees				Later semester enrollees			
	Training file 2017 and 2018	% SD	Test file 2019	% SD	Training file 2017 and 2018	% SD	Test file 2019	% SD
No tests taken	1348	51.93	771	44.2	1581	51.6	936	45.9
Only first test taken	1186	45.36	669	35.7	1337	42.8	794	36.3
First and second tests taken	1003	35.39	599	28.2	1056	27.6	652	22.4

Source: Own elaboration.

For each model, the main file was divided into two parts, one used to train the model and another used to validate it. It can be seen that the percentages of SD in the validation and training files

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

are similar, despite being samples from different years. The F1 metric was selected to compare the performance of the algorithms because it provides a good balance between precision and recall. To perform the analyses, the statistical software R version 4.2.1 and the Boruta packages were used for variable selection, and Caret and Tidy models for adjusting the predictive models and generating the Gini importance measure for each variable in the case of RF and XGB.

Description of the information analysis strategy

The sample consisted of data from those students who enrolled in the MAT001 course during the academic years ranging from the first semester of 2017 to the second semester of 2019. Since the objective is to predict SD in the course, these cases represent the available sample of all possible people who enrolled or could enroll in the course in the future.

The first step consisted of cleaning the data file provided by the Registration Department and the School of Mathematics. Then, the design of six models was proposed, which were evaluated by means of k-fold cross-validation (k=10).

The second step was to choose the variables for the models through a descriptive and correlational analysis in which variables with redundant information were excluded, such as the university admission grade, composed of the high school grade and the grade in the Academic Aptitude Test (AAT). For cases of variables that were highly correlated ($r > 0.80$), or that showed a linear relationship between them, only one was selected. Variables with an insufficient number of cases per category, such as nationality, were also excluded because they affected the estimates of the models.

The Boruta feature selection algorithm was then applied to select the variables with the greatest prediction potential for each model; thus, the total number of variables was reduced from 148 to 18 for first semester enrollees and to 22 for later semester enrollees.

As an example Figure 1 shows the selection made by Boruta for Model 1, which will be used to exemplify each phase of the procedure.

The data and code necessary to reproduce the other models can be consulted on the site https://acortar.link/qQNIyv

The third step, once the variables for each model were selected, consisted of using a 10-fold cross-validation to adjust the hyperparameters of the RF and XGB algorithms (LR does not have hyperparameters). For example, for the RF hyperparameters of the number of trees included in the model, the maximum depth of the tree and the minimum number of observations per node were adjusted, while for XGB the hyperparameters eta, gamma, max_dept and min_child_weight were adjusted. This procedure was applied for each algorithm of the six models.

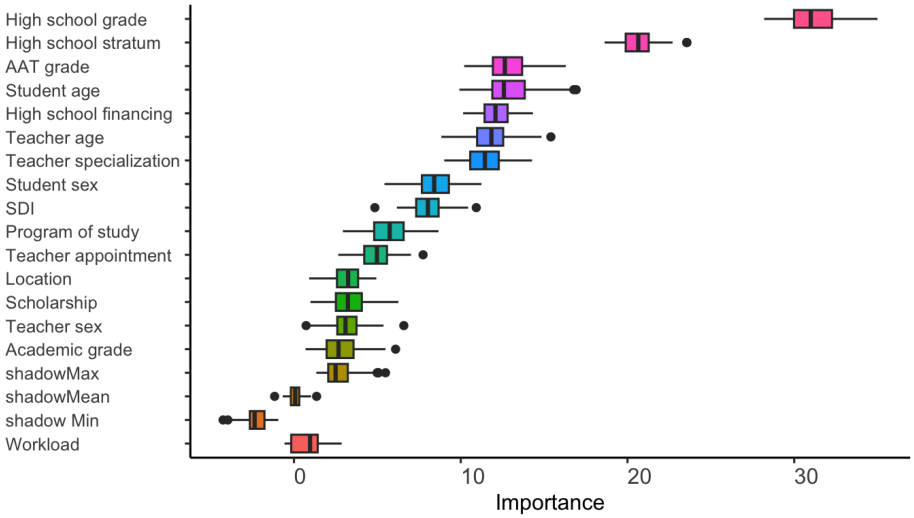
CIn a fourth step, once the optimal parameters for each algorithm in each model were determined, a new 10-fold cross-validation was performed to train the three algorithms in each of the six models, making sure to use a random seed that allowed the results to be reproduced and guaranteed that each

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

algorithm used the same fold in each of the 10 iterations. This was done with the training set (2017 and 2018 cohorts). As an example, Figure 2 shows the average of the comparison measure (F1 score) for each algorithm in model 1 (first semester enrollees, no course test grades).

Figure 1.

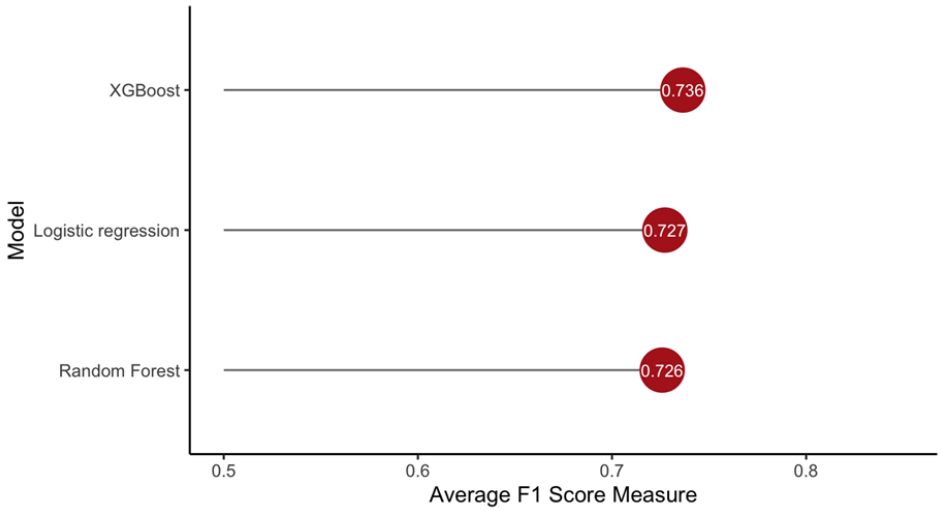
UNA: Variables selected by the Boruta algorithm for first semester enrollees without MAT001 test grades



Source: Own elaboration.

Figure 2.

UNA: Comparison of Logistic Regression, Random Forest and XGBoost algorithms by 10-fold cross-validation for first semester enrollees without MAT001 test grades



Source: Own elaboration.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

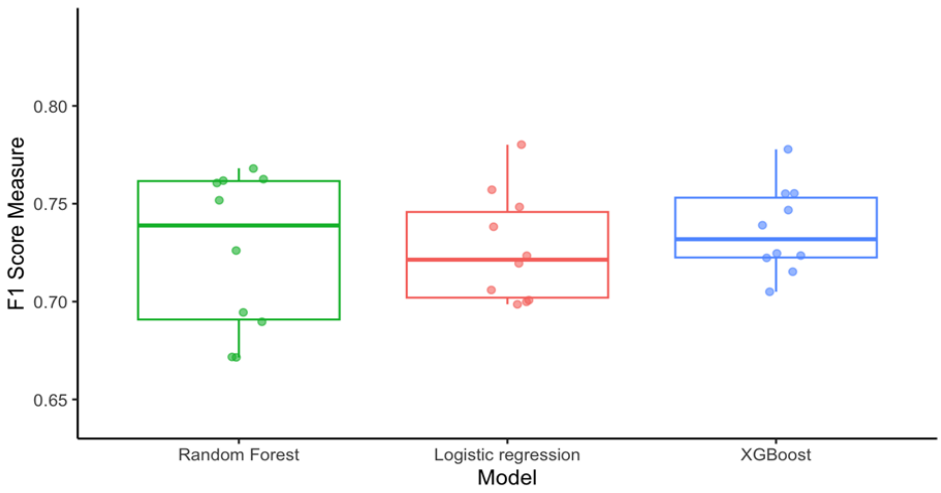
In the fifth step, a repeated measures ANOVA (each iteration is a replication), the results are evaluated to determine if there is evidence in favor of any algorithm in terms of predictive performance for each model. Before applying the ANOVA, the assumption of normality was verified with the Anderson Darling test, which provided p values greater than 0.05 in all cases. When applying the ANOVA, the null hypothesis of equal means was not rejected (p value >0.05), so there is no evidence in favor of any of the algorithms in the selected metric. As an example, Figure 3 shows the box diagram for the ANOVA of model 1.

In step 6, since there is no evidence that indicates any model has a significantly better performance than the other models according to the F1 metric, the three algorithms are evaluated on the corresponding test data (data never previously seen by the model) in the six models.

Finally, to determine the most relevant variables for the prediction of SD in each black-box algorithm, the Gini importance measure and the odds ratio measure were used in the case of logistic regression. Figure 4, Figure 5 and Figure 6 show the relative importance of each variable for each algorithm of Model 1.

Figure 3.

UNA: Box diagram representing the ANOVA that compares the Logistic Regression, Random Forest, and XGBoost algorithms by 10-fold cross-validation for first semester enrollees without MAT001 test grades

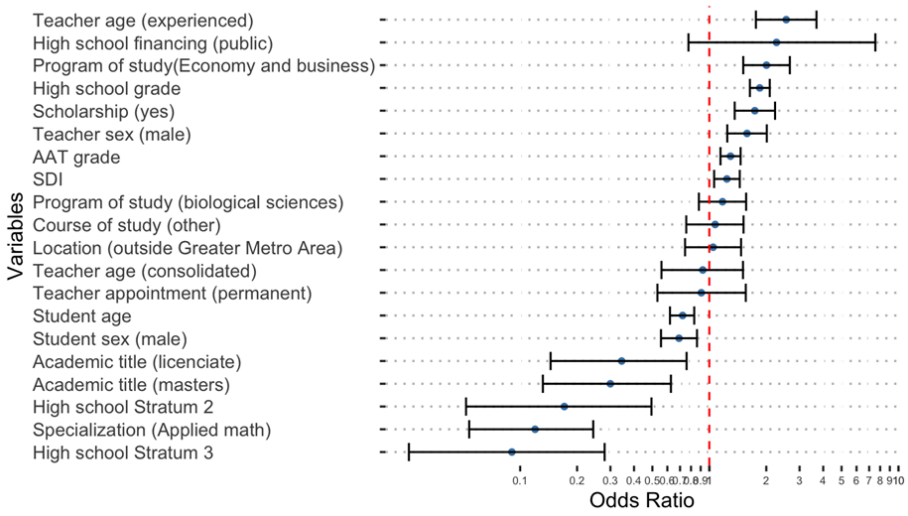


Source: Own elaboration.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Figure 4.

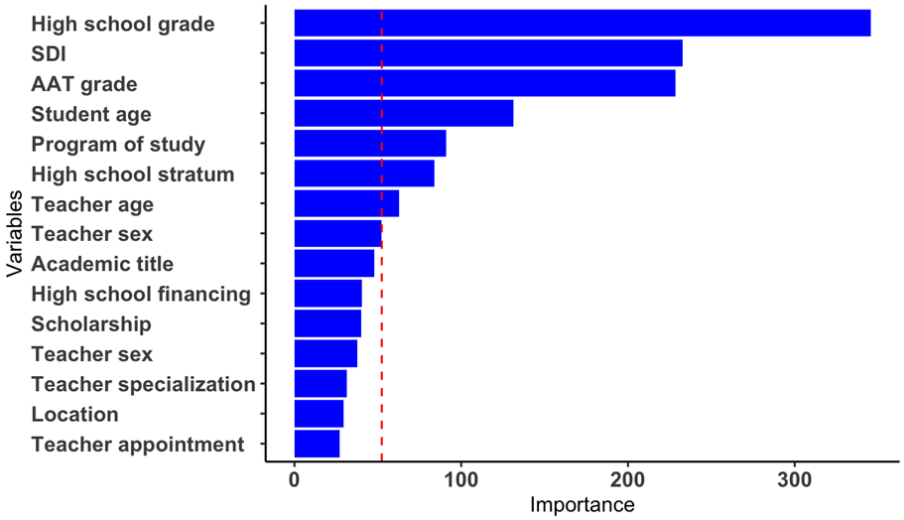
UNA: Importance of the variables according to the Logistic Regression algorithm for first semester enrollees without MAT001 test grades



Source: Zamora-Araya (2023b) p. 143.

Figure 5.

UNA: Importance of variables according to the Random Forest algorithm for first semester enrollees without MAT001 test grade

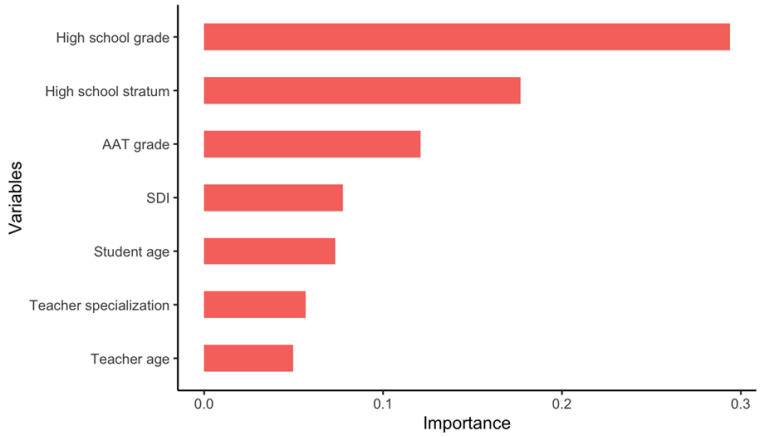


Source: Zamora-Araya (2023b) p. 144.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Figure 6.

UNA: importance of variables according to the XGBoost algorithm for first semester enrollees without MAT001 test grades



Source: Zamora-Araya (2023b) p. 144.

RESULTS AND DISCUSSION

One of the main results was the values of the F1 comparison metric obtained by evaluating the algorithms in the test files proposed in each of the models, which are summarized in Table 2.

As can be seen, the models have similar predictive power and the superiority of any of them cannot be affirmed. However, the Random Forest and the Logistic Regression showed slightly higher values in most of the models.

Table 2.

Value of the F1 metric for each model according to type of student and number of course tests in the year 2019

	Model according to included variables	Algorithm		
		Logistic regression	Random Forest	XGBoost
First semester enrollees	Only admission variables	0.62	0.61	0.63
	Includes first course test	0.70	0.70	0.65
	Includes first two course tests	0.71	0.71	0.70
Later semester enrollees	Only admission variables	0.66	0.68	0.67
	Includes first course test	0.73	0.72	0.65
	Includes first two course tests	0.70	0.70	0.68

Source: Own elaboration.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

The importance of the characteristics in each model varied, as can be seen in Table 3. For example, in the first model, the variables identified as relevant in the two algorithms with the best performance (LR and RF) were: high school grades and AAT grades, high school stratum, and age of both students and teachers; other less relevant variables were the SDI, teacher's degree, scholarship, the sex of the students and the teacher's specialty.

In several studies, the relationship between high school grades and AAT with the performance and permanence of students in the MAT001 course has been confirmed (Castillo-Sánchez et al., 2020; Zamora-Araya et al., 2020). Scholarships have also been shown to be a factor to decrease SD, as has the SDI (Zamora-Araya et al., 2020). Regarding the age of entry into the university and the sex of the students, the models identified that older age and being male increase the probability of SD, variables that are usually associated with gaps in previous educational levels or pressure to enter the labor market at an early age (Lázaro et al., 2020; Mora, 2016).

Table 3.

UNA: Variables considered relevant in each algorithm according to the proposed model

Model	Logistic Regression	Random Forest	XGBoost
1. No test grades (First semester enrollees)	Age (teacher) Program of study High school grade AAT grade Scholarship Sex (teacher) Age (student) Sex (student) Academic degree (teachers) High school stratum 2 High school stratum 3 Teacher specialization	High school grade SDI AAT grade Age (student) Program of study High school stratum Age (teacher) Sex (student)	High school grade High school stratum AAT grade SDI Age (student) Teacher specialization Age (teacher)
2. First test grade (First semester enrollees)	Test 1 Age (teacher) Program of study Scholarship High school grade Sex (student) Age (student) Academic degree (teachers) Teacher specialization	Test 1 High school grade Age (student) High school stratum SDI Teacher specialization	Test 1 High school grade Age (student)

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

3. First and second test grades (First semester enrollees)	Test 1 Test 2 Age (teacher) Teacher specialization	Test 1 Test 2 AAT grade SDI Age (student) High school stratum	Test 1 Test 2 SDI Age (student) Program of study
4. No test grade (Later semester enrollees)	Enrollment Scholarship High school grade AAT grade Weighted average Age (student) Sex (student) Appointment (teacher) Dropped subjects	Weighted average Failed subjects High school grade AAT grade Age (student) Enrollment Subjects passed Program of study	Subjects dropped Weighted average High school grade Age (student) Enrollment Appointment (teacher) Subjects passed Program of study Scholarship Sex (student)
5. First test grade (Later semester enrollees)	Test 1 Age (student) Dropped subjects Sex (student) Enrollment	Test 1 Failed subjects Weighted average High school grade Enrollment Age (student) AAT grade	Test 1 Weighted average Enrollment Subjects dropped Age (student) Sex (student) High school grade
6. First and second test grades (Later semester enrollees)	Test 1 Test 2 Failed subjects AAT grade	Test 1 Test 2 Weighted average AAT grade High school grade Failed subjects Age (student)	Test 1 Test 2 Weighted average Enrollment AAT grade Subjects dropped Age (student) Scholarship High school grade

Note. The names of the variables with the highest importance scores in each algorithm are highlighted in bold.

Source: Own elaboration.

The high school stratum is an important variable in the analysis of SD, since it is an indirect indicator of the socioeconomic conditions and educational opportunities of students in their high school

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

years. The results identified that students from Stratum 3, those with fewer educational opportunities, are the most prone to SD, which confirms the vulnerability of this student group (Rodríguez-Pineda y Zamora-Araya, 2014; Zamora-Araya et al., 2023).

In the case of students in the later semester enrollment category, academic variables such as high school grade, AAT grade and others from the academic history such as the number of failed subjects, weighted average (both from the previous semester) and the number of times they enrolled in the course (the greater the number, the greater the probability of SD) and to a lesser extent the age of the student and scholarships were also relevant. The test grades were the most important variables in the models with grades for the first two course tests.

In the case of institutional variables, receiving classes from experienced teachers (over 45 years old) and with a postgraduate degree in the area of mathematics education decreases the probability of SD compared to receiving courses from younger teachers or specialists in areas of applied and pure mathematics (see Figura 4). Esto comprueba la relevancia de la especialización en matemática educativa y la capacitación del profesorado en didáctica y evaluación, puesto que las tasas más altas de AE se presentan en los cursos iniciales (Munizaga et al., 2018; Opazo et al., 2021).

In summary, in most models high school grades, the grade on the AAT and the age of the students are important variables to be considered, regardless of whether the students enrolled in MAT001 in their first semester in the university or in later semesters, or whether or not the grades of the MAT001 tests are included in the models. This clearly indicates the importance of variables associated with the level of prior knowledge, not only in Mathematics but in other areas.

CONCLUSIONS

The purpose of this investigation was to identify relevant variables to assist in predicting the probability of SD of students who enrolled in the General Mathematics course, considering the type of student and the time at which SD occurs.

Academic variables were the most important in predicting SD, together with some characteristics of the teaching staff, the students' high school stratum and the school of origin of the students as well as their age of entry into the university and their sex.

For example, in the models for first semester enrollees in MAT001, both the student's high school stratum (material conditions and educational opportunities) and the teacher's specialization were significant in most of the algorithms, even when the grades of the course tests are incorporated. This clearly demonstrates that students from more vulnerable population segments have higher probabilities of SD, and highlights the importance of institutional programs and the assignment of professionals in the area of educational mathematics who design novel activities to improve academic results.

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Another example would be to use the model for first semester enrollees that does not include course tests to propose a preventive strategy to reduce SD before starting the school year. The model suggests that the most relevant variables are high school grades, the AAT grade, stratum, SDI and age of the student; based on this information, interventions can be designed that take these characteristics into account.

If the aim is to support students repeating the course before starting the school year, the analyses show that academic variables such as the weighted average grade, the number of subjects passed and dropped in the previous year, the age of the student and the characteristics of the teaching staff are relevant variables. These characteristics could be used to form student groups that can be attended by the appropriate teaching staff, according to the models, which would help to improve the levels of permanence and performance in the MAT001 course.

Given the importance of course test grades, in particular the first test grade, it is recommended that the faculty member with overall responsibility for the course propose evaluative and methodological actions that promote improved performance in these tests. These activities might include an increase in the number of tests and the possibility for students to make up the first test in the course, given the impact it has on the SD.

Likewise, future qualitative or mixed studies could further investigate the variables relevant to SD in Mathematics courses, such as the characteristics of the teaching staff or the degree of prior knowledge. They could also implement other algorithms, test those described here in other courses and explore the relationship with other constructs associated with SD such as student commitment, motivation, sense of institutional belonging, or the process of adjustment to university life. Based on the results of the present study, it is recommended to use the logistic regression algorithm, since its interpretation is simpler, making the relationship between SD and the independent variables in each model clearer to the users of the models.

REFERENCES

Aguilar-Fernández, E., Zamora-Araya, J. A. y Rodríguez-Pineda, M. (2024). Análisis de correspondencia simple para estudiar la relación entre factores del abandono escolar y el estrato del colegio de procedencia en la Universidad Nacional de Costa Rica [Simple correspondence analysis to study the relationship between factors of school dropout and stratum of school of origin at the Universidad Nacional de Costa Rica]. *Revista Educación, 48*(2), 1-20. https://doi.org/10.15517/revedu.v48i2.58519

Albreiki, B., Zaki, N. y Alashwal, H. (2021). Una revisión sistemática de la literatura sobre la predicción del rendimiento de los estudiantes utilizando técnicas de aprendizaje automático [A systematic review of the literature on the prediction of student performance using machine learning techniques]. *Ciencias de la Educación, 11*(9), 1-27. https://doi.org/10.3390/educsci11090552

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Alvarado, O. y Zambrano, S. M. (2020). *Modelo predictivo para determinar el fracaso de matemáticas en grado 11 usando machine learning [A model to predict mathematics failure in 11th grade using machine learning]*. [Degree project, Universidad Distrital Francisco José Caldas]. Repositorio Institucional Universidad Distrital - RIUD. https://repository.udistrital.edu.co/bitstream/handle/11349/25365/OmarAlvaradoSantosZambrano2020.pdf?sequence=1&isAllowed=y

Asamblea Legislativa. (2011). *Ley 8968 Protección de la persona frente al tratamiento de sus datos personales [Law 8968 Protection of persons against the processing of their personal data]*. http://www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=70975&nValor3=85989

Bäulke, L., Grunschel, C. y Dresel, M. (2022). Deserción estudiantil en la universidad: Una visión por fases sobre el abandono de los estudios y el cambio de carrera [Student dropout at the university: A view of phases of dropping out of studies and changing careers]. *Revista Europea de Psicología de la Educación, 37*(3), 853-876. https://doi.org/10.1007/s10212-021-00557-x

Behr, A., Giese, M., Teguim-Kamdjou, H. D. y Theune, K. (2020). Abandono de la universidad: una revisión de la literatura Dropping out of the university: A literature review]. *Revista de Educación, 8*(2), 614-652. https://doi.org/10.1002/rev3.3202

Cabrera, J. T., Álvarez, P. y González, M. (2006). El problema del abandono de los estudios universitarios [The problem of dropping out of university studies]. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa, 12*(2), 171-203. https://doi.org/10.7203/relieve.12.2.4226

Calva, K., Flores, M., Porras, H. y Cabezas-Martínez, A. (2021). Modelo de predicción del rendimiento académico para el curso de nivelación de la escuela politécnica nacional a partir de un modelo de aprendizaje supervisado [Academic performance prediction model for the leveling course of the national polytechnic school based on a supervised learning model]. *Latin-American Journal of Computing, 8*(2), 58-71. https://doi.org/10.5281/zenodo.5770905

Castaño, E., Gallón, S., Gómez, K. y Vásquez, J. (2004). Deserción estudiantil universitaria: Una aplicación de modelos de duración [University student dropout: An application of duration models]. *Lecturas de Economía,* (60), 39-65. https://doi.org/10.17533/udea.le.n60a2707

Castillo-Sánchez, M., Gamboa-Araya, R. y Hidalgo-Mora, R. (2020). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas [Factors that influence the dropout and failure of students in a university mathematics course]. *Uniciencia, 34*(1), 219-245. http://dx.doi.org/10.15359/ru.34-1.13

Guzmán, A., Barragán, S. y Cala Vitery, F. (2021). Deserción escolar en la educación superior rural: una revisión sistemática [Dropout in rural higher education: A systematic review]. *Fronteras de la educación, 6*, 1-14. https://doi.org/10.3389/feduc.2021.727833

Hernández, R., Fernández, C. y Batista, M. P. (2014). *Metodología de la investigación [Research methodology]* (6ta ed.). Mc Graw Hill.

Khan, A. y Ghosh, S. K. (2021). Análisis y predicción del rendimiento de los estudiantes en el aprendizaje en el aula: una revisión de los estudios de minería de datos educativos [Analy-

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

sis and prediction of student performance in classroom learning: A review of educational data mining studies]. *Educación y Tecnologías de la Información, 26*, 205-240. https://doi.org/10.1007/s10639-020-10230-3

Khoushehgir, F. y Sulaimany, S. (2023). Negative link prediction to reduce dropout in massive open online courses. *Education and Information Technologies,* 1-20. https://doi.org/10.1007/s10639-023-11597-9

Kilian, P., Loose, F. y Kelava, A. (2020). Predecir el éxito de los estudiantes de matemáticas en la fase inicial de la universidad con información dispersa utilizando enfoques de aprendizaje estadístico [Predicting the success of mathematics students in the initial phase of university with sparse information using statistical learning approaches]. *Fronteras de la Educación, 5*, 1-16. https://doi.org/10.3389/feduc.2020.502698

Lázaro, N., Callejas, Z. y Griol, D. (2020). Factores que inciden en la deserción estudiantil en carreras de perfil ingeniería informática [Factors that affect student dropout in computer engineering program courses]. *Revista Fuentes, 22*(1), 105-126. https://hdl.handle.net/11162/200868

López-Zambrano, J., Lara-Torralbo, J. A. y Romero-Morales, C. (2021). Predicción temprana del rendimiento del aprendizaje de los estudiantes a través de la minería de datos: una revisión sistemática [Early prediction of student learning performance through data mining: A systematic review]. *Psicotema, 33*(3), 456-465. https://reunido.uniovi.es/index.php/PST/article/view/17117

Mora, Y. (2016). *Estudio longitudinal de la deserción universitaria en el Instituto Tecnológico de Costa Rica [Longitudinal study of university dropout at the Instituto Tecnológico de Costa Rica]* [Master's thesis, Universidad de Costa Rica]. Repositorio SIBDI. https://repositorio.sibdi.ucr.ac.cr/handle/123456789/22144

Munizaga, F., Cifuentes, M. B. y Beltrán, A. (2018). Retención y abandono estudiantil en la educación superior universitaria en América Latina y el caribe: Una revisión sistemática [Student retention and dropout in higher education in Latin America and the Caribbean: A systematic review]. *Education Policy Analysis Archives, 26*(61), 1-36. http://dx.doi.org/10.14507/epaa.26.3348

Muñoz-Camacho, S. V., Gallardo, T., Muñoz-Bravo, M. y Muñoz-Bravo, C. A. (2018). Probabilidad de deserción estudiantil en cursos de matemáticas básicas en programas profesionales de la Universidad de los Andes-Venezuela [Probability of student dropout in basic mathematics courses in professional programs at the Universidad de los Andes-Venezuela]. *Formación Universitaria, 11*(4), 33-42. http://dx.doi.org/10.4067/S0718-50062018000400033

Opazo, D., Moreno, S., Álvarez-Miranda, E. y Pereira, J. (2021). Análisis de la deserción universitaria de primer año a través de modelos de aprendizaje automático: Una comparación entre universidades [Analysis of first-year university dropout through machine learning models: A comparison between universities]. *Matemáticas, 9*(20), 1-27. https://doi.org/10.3390/math9202599

Pascua-Cantero, P. M. (2016). Factores relacionados con la deserción en el primer y segundo año de estudio en la carrera de Enseñanza de la Matemática de la Universidad Nacional de Costa Rica [Factors related to dropout in the first and second year of study in the Mathematics Teaching degree program at the Universidad Nacional de Costa Rica]. *Revista Electrónica Educare, 20*(1), 96-118. http://dx.doi.org/10.15359/ree.20-1.5

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Proyecto ALFA-GUÍA. (2013). *Marco conceptual sobre el abandono. Hacia la gestión colectiva de un marco conceptual para analizar, predecir, evaluar y atender el abandono estudiantil en la educación superior [A conceptual framework for dropout. Towards the collective management of a conceptual framework to analyze, predict, evaluate and address student dropout in higher education]*. https://www.scribd.com/document/261888622/Marco-Conceptual-sobre-el-Abandono-pdf

Rodríguez-Pineda, M. y Zamora-Araya, J. A. (2014). *Análisis de la deserción en la Universidad Nacional desde una perspectiva longitudinal. Quinto informe estado de la educación [Analysis of dropout at the National University from a longitudinal perspective. Fifth Report on the State of Education].* Programa Estado de la Nación. https://doi.org/10.13140/RG.2.2.30416.66569

Román, M. (2017). Capítulo 5: La evolución de la educación superior [Chapter 5: The evolution of higher education]. In *Sixth Report on the State of Education* (pp. 241-308). Programa Estado de la Nación. https://hdl.handle.net/20.500.12337/1181

Shin, D. y Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education, 19*, 639-659. https://doi.org/10.1007/s10763-020-10085-7

Solís, M., Moreira, T., González, R., Fernández, T. y Hernández, M. (2018). Perspectives to predict dropout in university students with Machine Learning. *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 1-6. https://doi.org/10.1109/IWOBI.2018.8464191

Tinto, V. (1982). Definición de abandono escolar: una cuestión de perspectiva. *Nuevas Direcciones para la Investigación Institucional, 1982*(36), 3-15. https://doi.org/10.1002/ir.37019823603

Tinto, V. (1989). Definir la deserción: Una cuestión de perspectiva [Definition of school dropout: A matter of perspective]. *Revista de Educación Superior, 71*(18), 1-9. http://publicaciones.anuies.mx/pdfs/revista/Revista71_S1A3ES.pdf

Valencia, L. I., Guzmán, A. y Barragán, S. (2024). Deserción en programas de posgrado: un fenómeno poco explorado: una revisión exploratoria [Dropout in graduate programs: A poorly explored phenomenon – An exploratory review]. *Educación Cogent, 11*(1), 1-20. https://doi.org/10.1080/2331186X.2024.2326705

Wang, W., Zhao, Y., Wu, Y. J. y Goh, M. (2023). Factores de abandono de los MOOCs: Una revisión bibliométrica [Dropout factors from MOOCs: A bibliometric review]. *Biblioteca Hi Tech, 41*(2), 432-453. https://doi.org/10.1108/LHT-06-2022-0306

Xu, C., Zhu, G., Ye, J. y Shu, J. (2022). Minería de datos educativos: Predicción de abandono en los MOOCs de XuetangX [Educational data mining: Predicting dropout in XuetangX MOOCs]. *Cartas de Procesamiento Neuronal, 54*(4), 2885-2900. https://doi.org/10.1007/s11063-022-10745-5

Zamora-Araya, J. A. (2023a). *Modelo de un sistema de alerta temprana para reducir el abandono en el curso de Matemática General en la Universidad Nacional, Costa Rica* [*Model of an early warning system to reduce dropout in the General Mathematics course at the Universidad Nacional, Costa Rica]* [PhD thesis, Universidad Estatal a Distancia]. SIIDCA-CSUCA. https://catalogosiidca.csuca.org/Record/UNED.000099829

Revista Educación, 2025, 49(1), enero-junio ISSN: 0379-7082 / e-ISSN 2215-2644

Zamora-Araya, J. A. (2023b, noviembre). *Predicción del abandono temprano en estudiantes de nuevo ingreso en el curso de Matemática General utilizando algoritmos de aprendizaje supervisado [Predicting early dropout in new students in the General Mathematics course using supervised learning algorithms]* [Presentation].. Congreso Latinoamericano sobre Abandono en Educación Superior (CLABES), Temuco, Chile. https://clabes.uct.cl/wp-content/uploads/2024/06/Acta-XII-CLABES-Revision-final.pdf

Zamora-Araya, J. A., Aguilar-Fernández, E. y Rodríguez-Pineda, M. (2023). ¿Cuándo el abandono universitario se convierte en exclusión educativa? [When does university dropout become educational exclusion?]. *Revista Innovaciones Educativas, 25*(38), 97-115. http://dx.doi.org/10.22458/ie.v25i38.4212

Zamora-Araya, J. A., Gamboa, R., Hidalgo, R. y Castillo, M. (2020). Permanencia estudiantil en el curso de Matemática General de la Universidad Nacional, Costa Rica [Student retention in the General Mathematics course at the Universidad Nacional, Costa Rica]. *Actualidades Investigativas en Educación, 20*(1), 1-23. https://doi.org/10.15517/aie.v20i1.39815