Development of an intelligent system to predict university dropout rates in Colombia using machine learning techniques

Desarrollo de un sistema inteligente para predecir la deserción universitaria en Colombia mediante técnicas de machine learning

DOI: https://dx.doi.org/10.17981/ingecuc. 21.2.2025.08

Artículo de Investigación Científica.

Fecha de Recepción: 30/09/2025, Fecha de Aceptación: 14/10/2025

Ana Gabriela Banquez-Maturana
Universidad de la Salle. Bogotá, (Colombia)
abanquez56@unisalle.edu.co

Ángel Manuel Benavides-González
Universidad de la Salle. Bogotá, (Colombia)
abenavides80@unisalle.edu.co

Juan David Rodríguez-Cerón
Universidad de la Salle. Bogotá, (Colombia)
jurodriguez61@unisalle.edu.co

Heriberto Alexander Felizzola-Jimenez
Universidad de la Salle. Bogotá, (Colombia)
healfelizzola@unisalle.edu.co

To cite this paper
A. Banquez-Maturana, J. Rodríguez-Cerón, A. Benavides-González & H. Felizzola- Jimenez “Development of an intelligent system to predict university dropout rates in Colombia using machine learning techniques,” *INGE CUC*, vol. 21, no. 2, 2025. DOI: https://dx.doi.org/10.17981/ingecuc.21.2.2025.08.

Abstract

Introduction: Student dropout is a critical global challenge with profound socioeconomic and institutional impacts. On average, one in five students drop out of school, limiting social mobility, deepening inequalities, and reducing the sustainability of education systems.
Objective: Student dropout is a critical global challenge with profound socioeconomic and institutional impacts. On average, one in five students drop out of school, limiting social mobility, deepening inequalities, and reducing the sustainability of education systems.
Method: The study, framed within Design Science Research (DSR), used a longitudinal dataset of 104,147 records from a Colombian university. Rigorous preprocessing was applied, including reclassification of the target variable and engineering of 27 predictor features. Seven algorithms were evaluated, selecting LightGBM, which was optimized in its hyperparameters and balanced with SMOTE (Synthetic Minority Over-sampling Technique).
Results: LightGBM proved to be the superior algorithm with a weighted F1-Score of 0.8125. The optimized model achieved an overall accuracy of 87% and an F1-Score of 0.83 for the “Dropout” class. Strategic calibration of the decision threshold to 0.45 raised the recall to 87%, correctly identifying 1,447 of 1,654 actual dropouts. SHAP analysis confirmed that REAL_PROGRESS_PERCENTAGE was the most influential predictor with an impact of 1.45.
Conclusions: Cumulative academic performance, grade trends, and actual progress percentage are the most decisive predictors of dropout, in interaction with socioeconomic variables such as income stratum and demographic variables such as age group.

Keywords: University dropout; Applied artificial intelligence; Machine learning; Deep learning; SHAP values; Dropout prediction; Colombia.

Resumen

Introducción: La deserción estudiantil es un desafío crítico a nivel global con profundos impactos socioeconómicos e institucionales. En promedio, uno de cada cinco estudiantes abandona sus estudios, situación que limita la movilidad social, profundiza inequidades y reduce la sostenibilidad de los sistemas educativos.
Objetivo: Desarrollar un sistema inteligente para predecir la deserción universitaria en Colombia teniendo en cuenta el riesgo de abandono académico, con el fin de implementar intervenciones tempranas y personalizadas.
Metodología: El estudio, enmarcado en la Ciencia del Diseño (DSR), utilizó un dataset longitudinal de 104,147 registros de una universidad colombiana. Se aplicó un riguroso preprocesamiento, incluyendo la reclasificación de la variable objetivo y la ingeniería de 27 características predictoras. Se evaluaron 7 algoritmos, seleccionando LightGBM, el cual fue optimizado en sus hiperparámetros y balanceado con SMOTE (Synthetic Minority Over-sampling Technique).
Resultados: LightGBM demostró ser el algoritmo superior con un F1-Score ponderado de 0.8125. El modelo optimizado alcanzó una precisión global del 87% y un F1-Score de 0.83 para la clase “Desertó”. La calibración estratégica del umbral de decisión a 0.45 elevó el recall al 87%, identificando correctamente a 1,447 de 1,654 desertores reales. El análisis SHAP confirmó que el PORCENTAJE_AVANCE_REAL fue el predictor más influyente con un impacto de 1.45.
Conclusiones: El rendimiento académico acumulado, la tendencia de las calificaciones y el porcentaje de avance real constituyen los predictores más determinantes del abandono, en interacción con variables socioeconómicas como el estrato de ingreso y demográficas como el grupo etario.

Palabras clave: Deserción universitaria; Inteligencia artificial aplicada; Machine Learning; Deep learning; SHAP values; Predicción de la deserción; Colombia.

Barranquilla. ISSN 0122-6517 Impreso, ISSN 2382-4700 Online

INTRODUCTION

Student dropout in higher education is a critical global challenge with profound social, economic, and institutional impacts. International studies have shown that on average, about 20% of students who begin university studies do not complete them [1], [2], while in Latin America, only half of enrolled young people graduate [3]. This situation not only implies the loss of human capital and the investment made in the training of each student, but also limits social mobility, deepens inequality, increases performance rates, generates fewer job opportunities for vulnerable groups, reduces national productivity, and reduces the sustainability of educational systems [4], [5]. At the individual level, dropout is associated with a mix of socioeconomic, academic, personal, sociocultural, and emotional factors [6], [7]. Therefore, anticipating and understanding this phenomenon is essential for higher education institutions, with the support of government entities, to design effective prevention strategies instead of simply reacting once the dropout has already occurred [8].

Understanding and addressing university dropout rates has led to various approaches to the problem over the last few decades [9], [10], [11]. These include statistical analysis and monitoring using socioeconomic, academic and institutional indicators that have made it possible to identify and study the factors associated with student dropout [12]. Moreover, some studies have relied on surveys and interviews to explore students' motivations for dropping out [12], [13]. However, these approaches have significant limitations, such as the multifactorial complexity of the phenomenon, the limited capacity to anticipate specific dropout cases early, and the difficulty in dealing with large volumes of heterogeneous information [14]. These restrictions have driven the need to resort to more advanced analytical tools capable of integrating information from various sources and types, and generating more accurate predictions that allow higher education institutions to intervene in a timely manner through effective prevention mechanisms [15].

In recent years, advances in data analytics, machine learning, and artificial intelligence have proven useful in addressing the prediction of university dropout risk [16]. Supervised learning models allow students to be classified into different risk levels, which facilitates the construction of early warning systems that support institutions in making timely and evidence-based decisions [15]. These systems have become a fundamental tool to optimize institutional management since they facilitate the targeting of resources and the design of support programs for students at risk of dropping out [17], [18]. However, predicting who will drop out is not enough for effective management, since it is equally important to interpret and explain the factors that generate such risk so that prevention actions are based on a clear understanding of the underlying causes, which allows the design of interventions tailored to the profile and context of each student [19], [20].

One of the main challenges in the application of machine learning models to university dropout prediction is their black-box nature [21]. When algorithms generate predictions without offering a clear explanation of how the factors that contribute to this prediction are, it is a limitation in their implementation in early warning systems because it makes it difficult to design intervention strategies [22]. To overcome this limitation, interpretability approaches such as SHAP values have emerged, which allow to break down the prediction of an algorithm and explain the individual contribution of each variable to the risk of dropping out [23]. This interpretability capacity offers significant advantages, since it makes it possible to identify the risk factors associated with each student and consequently the design of personalized and effective intervention strategies to prevent academic dropout [17].

The purpose of this research was to develop an intelligent system for predicting university dropout rates in Colombia, based on the analysis of academic dropout risk and with the intention of facilitating early and personalized interventions. To this end, we analyzed the factors associated with university dropout rates in Colombia, identifying relevant academic, socioeconomic, and demographic variables through a case study. We designed an artificial intelligence-based predictive model capable of detecting students at risk of dropping out. We also evaluated its effectiveness and accuracy using standard tests and metrics, making the necessary adjustments to optimize its performance in the Colombian university context.

In this regard, the following questions arise: What are the academic, socioeconomic, and demographic factors that influence university dropout rates in Colombia? How can an artificial intelligence-based predictive model be developed to identify students at risk of dropping out using advanced machine learning techniques? How can the effectiveness and accuracy of the proposed predictive model be evaluated using standard metrics, ensuring its fit and applicability in the Colombian university context?

The study, framed within the Design Science Research (DSR) approach, used a longitudinal dataset of 104,147 records belonging to a Colombian public university. The methodology included a rigorous data preprocessing process, which included the reclassification of the target variable and the engineering of 27 predictor features. Seven machine learning algorithms were evaluated, and LightGBM was selected as the most appropriate. This model was optimized. in its hyperparameters and incorporated the synthetic minority oversampling technique (SMOTE) to balance classes and improve the system's predictive capacity. Additionally, the SHAP values explainability method was used to identify the determining factors that influence the risk of dropping out, providing valuable information to guide mitigation actions. In this way, the article contributes to the construction of more accurate early warning systems that, in turn, support higher education institutions in implementing timely, relevant, and effective interventions to combat student dropout.

LITERARY REVIEW

University dropout is understood as the phenomenon in which students interrupt their studies before completing them [8], [19], [24], [25], [26]. Although it is present at different educational levels, it is of greater relevance and concern in higher education institutions [27], [28]. The literature defines it as a complex, multicausal and dynamic process, in addition to constituting a global challenge [7], [29], [30], [31]. According to the parameters reviewed in the state of the art, we seek to demonstrate both the factors associated with this phenomenon, through relevant causal variables, and the artificial intelligence techniques used for its prediction, which is recognized as a serious socioeconomic problem [32]. At the individual level, it limits the personal and professional development of students [33], while at the institutional level it causes financial losses, hinders organizational growth and affects educational quality [2] [34].

The central objective of numerous investigations has been to anticipate dropout as early as possible in order to design timely interventions that improve retention rates and raise educational quality [35], [36], [37], [38], [39], [40], [41], [42]. The causes that explain it are diverse and include personal, economic, academic, institutional and sociodemographic factors [26]. Among the personal factors, depression or persistent academic failure have been identified [43], [44]. At the economic level, financial limitations are a key determinant [43], [45], [46]. At the academic level, previous performance and the results obtained in the first semester have an influence [36], [47]. At the institutional level, the conditions of support and integration into the academic system intervene [26]. From a social and sociodemographic perspective, variables such as marital status, gender, or cultural context also influence student retention [3], [22], [24], [31], [32]. In Colombia, multiple studies have delved into these dimensions, identifying relevant academic, socioeconomic, and demographic variables from case studies, as presented in Table 1.

In this scenario, the prediction of student dropout in higher education has been consolidated as an active field of research, supported using Machine Learning (ML) techniques and algorithms (Table 2). Among the most recurrent algorithms is Logistic Regression, recognized for its interpretation capacity ([1], [2], [3], [15], [43], [34], [48], [32]. It has been reported that this model achieves predictability levels of up to 69.9% [1] and that, in certain cases, it outperforms techniques such as Naive Bayes, Neural Networks, Decision Trees, Support Vector Machine and Random Forest [6]. In a predictive analysis, the gender variable turned out to be the only statistically significant under this approach, reaching a classification accuracy of 0.881 [49].

Neural Networks, both artificial and deep, also exhibit high predictive potential [2], [32], [50], especially in virtual learning contexts [51], [52]. A hybrid model that combined Logistic Regression and Neural Networks obtained an accuracy of 96% [32]. Likewise, an Artificial Neural Network (ANN) model reached an accuracy of 90.3%, in which the vulnerability index stood out as the most significant variable [49]. In another experiment, a six-layer model achieved an accuracy of 98.97%, surpassing XGBoost, which registered 87.1% [53]. Additional reports indicate that DNN and CNN-LSTM models exceed 90% accuracy [52] and that neural network-based systems reach values close to 92% [54]. The review of [55] shows that Neural Networks have been applied in 50% of the reviewed studies, although other comparisons point out that LightGBM can obtain better metrics, such as the F1-score [56].

Decision Trees stand out for their interpretive simplicity [46]. A model applied to adaptability in online education achieved an accuracy of 92% [8], while another system achieved 91% with a minimum error rate of 9% [57]. Regarding the Random Forest algorithm, several investigations have shown its outstanding performance: in virtual environments it achieved the best prediction results [58]; in combination with genetic algorithms, it achieved an average accuracy of 93.11% [59]; and in other studies, it consistently exceeded 90% accuracy [60], [61].

In the case of XGBoost, an accuracy of up to 90.3% has been documented in models aimed at at-risk students [5], [62], and a superior performance with AUC metrics of 91.3% in relation to other algorithms [63]. Although in certain scenarios some neural network models have managed to outperform XGBoost, the latter remains one of the most consistent algorithms in the specialized literature [53]. LightGBM, on the other hand, has achieved outstanding metrics, with an F1-score of 84% and superior results compared to other algorithms in direct comparisons [56], [64].

The K-Nearest Neighbors algorithm has reported an accuracy of 91% when integrating academic and socioeconomic variables in the dropout prediction [65]. Similarly, ensemble models have shown superior performance when combining different techniques and optimizing the classification of at-risk students. One such model, which integrated Logistic Regression, Neural Networks and Decision Trees, successfully classified 89% of cases and accurately identified 98.1% of dropouts [2]. Likewise, a stacking ensemble that integrated Random Forest, XGBoost, Gradient Boosting and Neural Networks reported improved performance in the dropout prediction [66].

For its part, approaches such as the Analytic Hierarchy Process, reached predictability levels of 64.6% [1]. There is no universally superior algorithm, since its performance depends on the characteristics of the dataset, the type of problem and the evaluation metrics used. However, algorithms such as Random Forest, XGBoost and LightGBM, together with ensemble models and neural networks, stand out in the literature for their high levels of accuracy. Logistic Regression, although in some contexts it shows slightly lower performance, retains significant value for its interpretability, which is relevant when seeking to understand the reasons that explain dropout and not just predict it.

TABLE 1. FACTORS ASSOCIATED WITH UNIVERSITY DROPOUT IN COLOMBIA

No	General Factors	Relevant variables	Number of items	Articles
1	Socioeconomic and Economic Factors	Debts; tuition arrears; scholarship status; scholarship type and percentage; scholarship loss; financial aid received; expected family contribution; income; family income; financial capacity; tuition costs; payment plans; financing and percentage; stratum; financial condition; occupation; decent employment; unemployment rate; inflation rate; GDP; parental financial support; extracurricular activities; geographic barriers; readmission costs.	26	[1], [2], [3], [4], [5], [31], [32], [42], [44], [54], [56], [57], [59], [63], [64], [66], [29], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76]
2	Academic Factors	Academic performance; grades; course completion; credits enrolled; GPA; prior education and previous degrees; academic integration and adaptation; degree satisfaction; career guidance; study methods; admissions processes; class attendance and participation; time spent studying; learning habits; performance in the first semesters.	31	[2], [3], [4], [5], [6], [32], [42], [44], [47], [50], [51], [54], [57], [58], [59], [60], [61], [63], [64], [31], [68], [69], [70], [71], [72], [77] [38], [78], [79]
3	Personal, Emotional and Health Factors	Marital status; gender; ethnicity; special educational needs; displacement status; individual motivation; physical and mental health; depression; stress; resilience; personal and family problems; pregnancy and family responsibilities; living with parents; alcohol use; maladjustment; bullying; need for emotional support; time management; dealing with emotions and frustrations; attention span and level of commitment.	22	[1], [2], [3], [4], [5], [8], [32], [42], [56], [57], [59], [63], [66], [19], [31], [68], [69], [70], [72], [73], [75], [76], [77]
4	Institutional and Pedagogical Factors	Tuition payment; support for students with special needs; course; schedule; pedagogical proposal; unsatisfactory methodologies; institutional dissatisfaction; institutional policies; year of enrollment; faculty; location; modality; pedagogical modification; scholarships; campus; teaching quality; teaching; exams; payment plans; curricula; monitoring; guidance; counseling; satisfaction with content; learning media; academic system.	22	[1], [2], [3], [4], [5], [31], [32], [42], [56], [57], [59], [63], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77]
5	Demographic Factors	Gender; age; marital status; nationality; place of origin; region of origin; change of residence; area of residence; province; postal code; occupation; displaced person; ethnicity; family type; number of children; number of family members; legal guardian; school of origin; type of school; profile of school of origin; level of education completed; type of admission; sociodemographic information.	20	[2], [3], [4], [5], [31], [32], [42], [47], [54], [56], [57], [59], [63], [68], [69], [70], [71], [38], [77]
6	Macrosocial Context Factors	Market demands; educational inequality; economic inequality; investment in higher education; political conflicts; social movements; social integration; social vulnerability; accessibility to higher education; globalization; family obligations; regional needs; social development; COVID-19 pandemic; technological change; poverty; corruption; state policies; educational sustainability.	21	[2], [3], [4], [5], [31], [32], [42], [57], [59], [63], [68], [69], [70], [72], [74]
7	Technological Factors and Online Education	E-learning environments; MOOCs; online platforms; blended and distance learning; open education; digital activity level and engagement; online class attendance history; internet access; devices used; LMS logs on access days, number of sessions, and connection times; online self-study; study habits and time spent on virtual courses.	18	[2], [5], [6], [8], [29], [31], [42], [57], [59], [63], [67], [68], [69], [70], [72], [77], [80], [81]
8	Relational and Coexistence Factors	Family and support relationships; peer problems; social support; recommendations from friends or family; social integration; family support; support networks; university environment; student quality and effort; bullying; relationships with teachers and peers; living with parents; legal guardian; communication with family; emotional relationships.	13	[1], [2], [31], [32], [42], [57], [59], [63], [68], [69], [70], [72], [74], [77]

Source: Authors, 2025.

TABLE 2. LIST OF ARTIFICIAL INTELLIGENCE (AI) AND PREDICTIVE ANALYSIS TECHNIQUES USED IN THE IDENTIFICATION AND PREVENTION OF UNIVERSITY DROPOUTS

No	Authors	Applied ML Techniques/ Accuracy	Explainability techniques used
1	[32]	HLRNN: 96%; LR: 94%; ANN: 93%; MLP: 92%; NB: 85%; RF: 93%; XGB: 93%; SVM: 92%	SHAP, LIME
2	[4]	RF: 80.56%; XGB: 78.98%	-
3	[2]	LR: 96.7%; ANN: 97.1%; CART-D: 96.6%; CART-N: 96.2%; ENS: 97.3%; RF: 97.6%	-
4	[3]	RF: 0.810 ± 0.053; XGB: 0.823 ± 0.042; LGBM: 0.813 ± 0.044; CAT: 0.812 ± 0.042	LIME
5	[31]	MLR: 76.44%	-
6	[42]	RF: 85%; FTT: 87%	SHAP
7	[77]	GCA-NN: 98%	-
8	[68]	AdaB: 92.7%; XGB-B: 93.9%; XGB-MC: 86.4%	SHAP
9	[69]	GINI: 92.20%; AUC: 96.10%	-
10	[59]	RF: 86.79%; RF+GA: 92.84%	-
11	[5]	XGB: 90.3%	-
12	[63]	XGB+IFS: 91%	-
13	[70]	LR-FM: 82.2%; LR-SBM: 82.9%; XGB: 83.3%; LGBM: 83.5%	-
14	[71]	LR: 96.41%	-
15	[72]	XGB: 83.3%; LGBM: 83.5%	-
16	[57]	K-M: 44.3%; LR: 35.1%; DT: 90.5%	-
17	[64]	DNN: 79.8%; LGBM: 82.6%	-
18	[19]	DT: <74%; LR: <74%; RF: 91.90%; AdaB: 95.20%; XGB: 95.20%	-
19	[6]	LR: 90.34%; DT: 89.32%; SVM: 89.72%; RF: 90.94%; MLP: 89.77%; NB: 90.84%	-
20	[29]	RNN: 53%; RF-ent: 65%; DFFNN: 63%; RF: 65%; AML: 67.4%; ANN-LSTM: 82%	-
21	[8]	DT: 92%	-
22	[44]	XGB: 99.75%	-
23	[47]	LR: 99.58%; ANN: 97.72%; DT: 99.53%	-
24	[38]	LGBM: 79.6%	SHAP
25	[80]	SVRQ: 87.16%	-
26	[51]	LM-BP: 87.36%; BFGS-BP: 88.9%; SGD: 79.74%; RBFNN: 89.02%; DT: 81.61%; BN: 78.79%; SVM: 87.36%; NNC: 90.19%	-
27	[56]	DNN: 79.8%; LGBM: 82.6%	-
28	[60]	RF: 90%; GBT: 93.6%	-
29	[78]	LR: 62.26%	-
30	[73]	ANN: 87.0%; GB: 63.8%; ENS: 60.6%; SDP: 81.9%	-
31	[61]	DKT: 81.1%; DKVMN: 79.7%; SAKT: 80.1%; DeepFM: 83.3%; DKT: 73.0%	-
32	[54]	ANN: 92%	-
33	[50]	ANN: 81%; SVM: 84%; GB: 76%; RF: 82%; KNN: 88%	-
34	[65]	KNN: 91%	-
35	[67]	ENS: 98.52%	-
36	[74]	MLP: 97%	XAI
37	[66]	FNN: 76.67%; RF: 91.66%; GB: 86.66%; XGB: 91.66%; ENS: 92.18%	-
38	[75]	NB: 76%; DT: 96%; RF: 96%; ANN: 81%	-
39	[43]	GB: ~72%; RF: ~82%; SVM: ~80%; ENS: 91.5%	-
40	[81]	LR: 86.79%; DT: 82.86%; BAG: 82.19%; RF: 85.51%	-
41	[58]	DT: 91%; RF: 96%; SVM: 81%; DNN: 85%	-
42	[79]	DT: 93%; NB: 88%; RF: 93%; LR: 94%	-
43	[76]	DT: 91%; MLP: 98%	-

Source: Authors, 2025.

METHODOLOGY

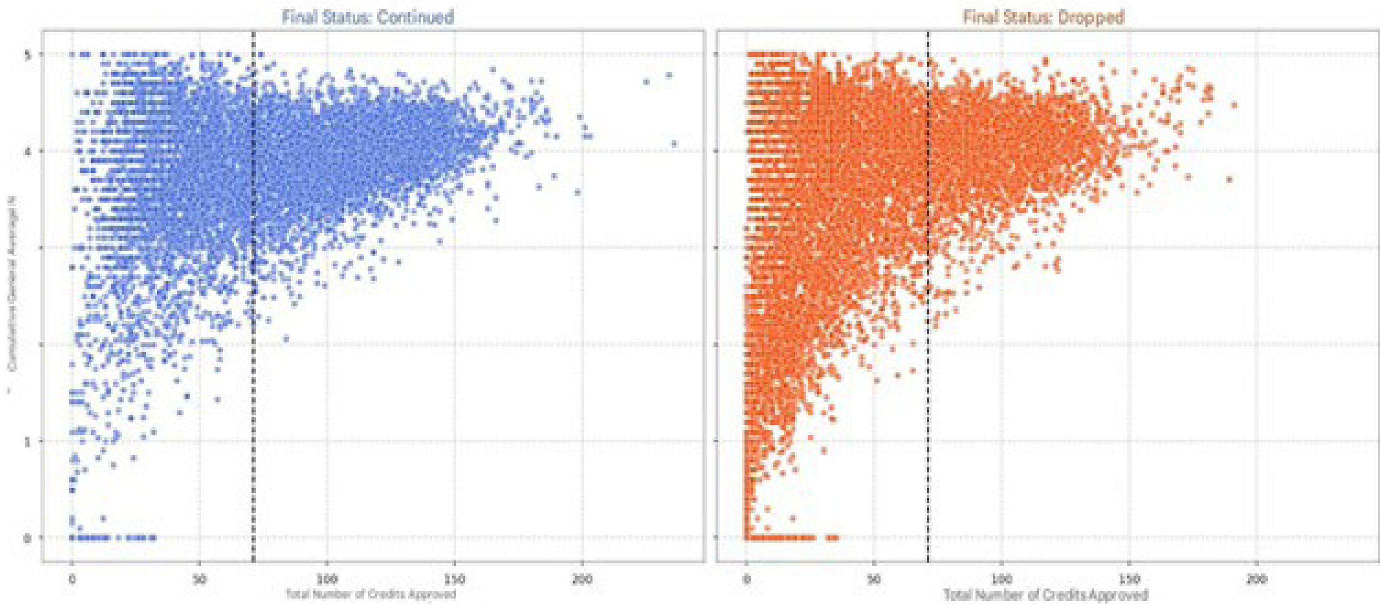
This chapter details the systematic process followed to develop the intelligent system, an artifact built under the Design Science Research (DSR) paradigm.

Data Source and Initial Diagnosis

The research used as a primary source an institutional dataset of 104,147 longitudinal records from a Colombian university, completely anonymized in accordance with Law 1581 of 2012 and the FAIR principles.

During the preliminary exploratory analysis (EDA), a diagnosis of the original target variable was performed. An atypical dropout rate of 86.2% was detected, suggesting possible label contamination. Figure 1 presents visual evidence of this anomaly.

Fig 1. Evidence of target variable contamination. The profiles of a subgroup of "Dropouts" (right) are indistinguishable from the high-performing "Continuers" (left), suggesting they correspond to mislabeled graduates. The dotted line indicates a common credit threshold for graduation.



Source: Authors, 2025.

Analysis of the figure reveals an overlap in profiles between a subgroup of "dropouts" and high-performing students, validating the hypothesis of systemic contamination in which graduates were being incorrectly grouped.

Preprocessing and Feature Engineering

To correct this fundamental inconsistency, a preprocessing pipeline was implemented, the core of which was a supervised reclassification protocol. This protocol reclassified each student into one of three mutually exclusive and properly defined categories: Continued, Actual Dropout, or Graduated. The result was a validated dataset of 22,275 records. On this clean base, a feature engineering process was applied to transform raw variables into 27 semantic predictors. This process included both the discretization of continuous variables (e.g., ACADEMIC_LEAGUE) and the creation of new interaction variables to model student trajectory (e.g., ACTUAL_PROGRESS_PERCENTAGE). Finally, numerical variables were standardized (StandardScaler) and categorical variables were encoded (OneHotEncoder). Table 3 details the final set of predictors.

TABLE 3. FINAL VARIABLES FOR THE PREDICTIVE MODEL

Variable	Description
Group I: Categorical Characteristics	
ACADEMIC_LEAGUE	Cumulative academic performance level
PROGRESS_MILESTONE	Career progress by accumulated credits
RACE_STAGE	Stage of the degree by semesters completed
FINAL_PERFORMANCE_LEVEL	Performance level in the last semester
ACADEMIC_LOAD_LEVEL	Commitment level (credits enrolled)
EFFICIENCY_LEVEL	Pass rate in the last semester
SEMI-ANNUAL_FAILURE_LEVEL	Number of subjects failed in the last semester

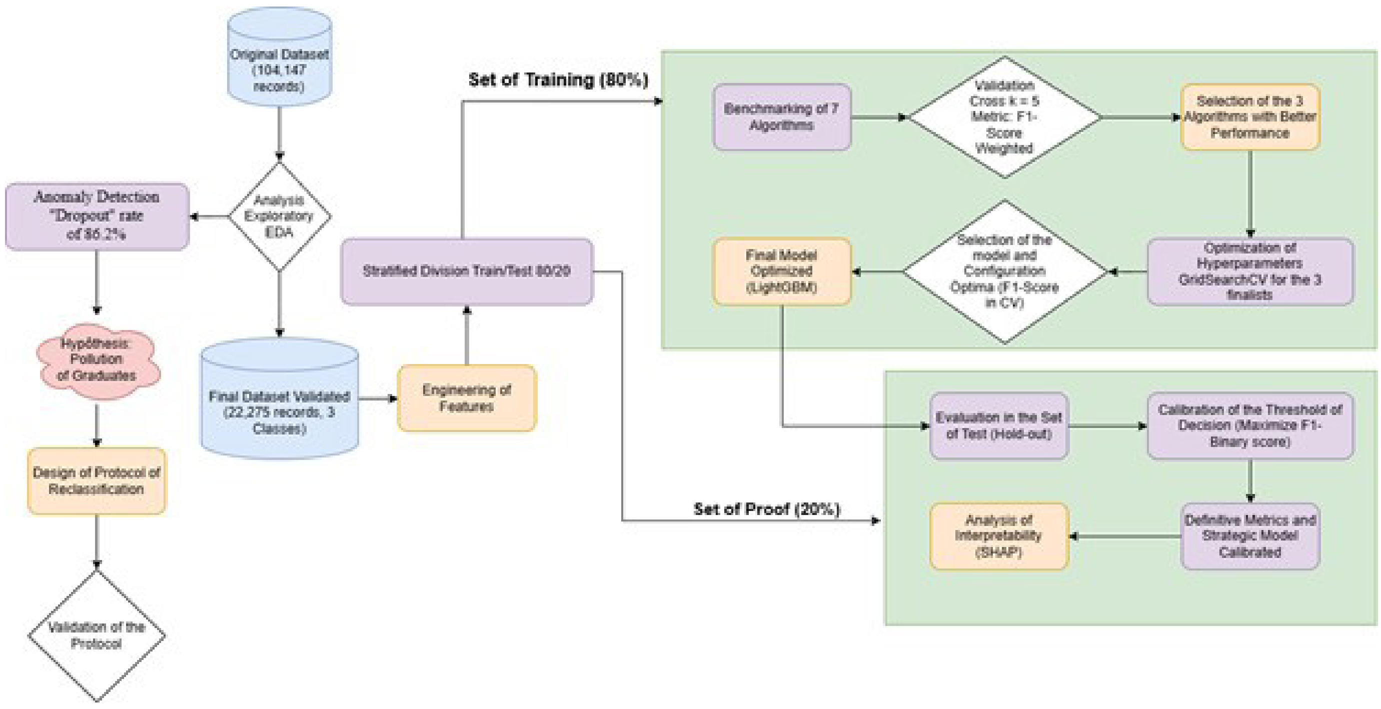
Variable	Description
INCOME_STRATUM	Student’s family income level
AGE_GROUP	Student age group
FAMILY_SUPPORT_STRUCTURE	Marital status and family structure of the student
PROGRAM	Student-specific academic program
Group II: Binary Characteristics	
HAS_SCORE	Proxy for Undergraduate (1) or Graduate (0) students
CAREER_CHANGE	Indicator of whether the student has changed careers
USA_RESOURCES	Indicator of whether the student uses resources
SUPPORT_TYPE	Indicator of whether you receive financial support or a scholarship
Group III: Numerical Characteristics	
CUMULATIVE_OVERALL_AVERAGE	Student’s cumulative numerical average
AVERAGE_GRADE_GRADES	Numerical average of the last recorded semester
NUMBER_OF_SEMESTERS_TAKEN	Number of semesters the student has completed
LAST_SEM_APPROVAL_RATE	Numerical pass rate for the last semester
PROGRAM_LENGTH	Official duration in semesters of the program
PERCENTAGE_ACTUAL_PROGRESS	Relative progress in the degree (semesters/duration)
ACCUMULATED_FAILURE_RATE	Rate of failed subjects per semester taken
PERFORMANCE TREND	Difference between recent and historical average
ACADEMIC_COMMITMENT_ WEIGHTED	Recent efficiency-weighted cumulative average

Source: Authors, 2025.

Experimental Design and Modeling

Once the final data set was prepared, the experimental design for the development of the predictive model was carried out, whose flowchart is presented in Figure 2.

Fig 2. Flowchart of the model training, optimization and evaluation pipeline.


Original Dataset (104,147 records)
Analysis Exploratory EDA
Anomaly Detection "Dropout" rate of 86.2%
Hypothesis: Pollution of Graduates
Design of Protocol of Reclassification
Validation of the Protocol
Final Dataset Validated (22,275 records, 3 Classes)
Engineering of Features
Stratified Division Train/Test 80/20
Set of Training (80%)
Set of Proof (20%)
Benchmarking of 7 Algorithms
Validation Cross k = 5 Metric: F1-Score Weighted
Selection of the 3 Algorithms with Better Performance
Optimization of Hyperparameters GridSearchCV for the 3 finalists
Selection of the model and Configuration Optima (F1-Score in CV)
Final Model Optimized (LightGBM)
Evaluation in the Set of Test (Hold-out)
Calibration of the Threshold of Decision (Maximize F1-Binary score)
Definitive Metrics and Strategic Model Calibrated
Analysis of Interpretability (SHAP)

Source: Authors, 2025.

To build the model, we followed a clear three-stage process implemented in Python and Scikit-learn. First, we searched for the best algorithm type. To do this, we put seven different models through their paces in a cross-validation evaluation (k=5), judging them with the Weighted F1-Score. From this competition, three finalists qualified. In the second stage, we focused on fine-tuning these finalists. We applied a GridSearchCV search to find the best combination of hyperparameters for each model, a process in which LightGBM proved to be the most robust. The final stage was the acid test: we took the optimized LightGBM and evaluated it against the test dataset, which we had kept intact. This final exam was comprehensive, covering not only technical metrics such as precision and recall, but also a strategic calibration to create a “Strategic Model” and an interpretability analysis with SHAP. All development was done in Google Colaboratory, leveraging GPU acceleration and key libraries such as Pandas, Imbalanced-learn, and the boosting models themselves.

RESULTS

Model Selection and Optimization

The first step in determining the final predictive model was a thorough benchmarking. The performance of seven different algorithms was measured to identify the most robust, and the results (Table 4) highlighted a clear trend: Gradient Boosting models were superior. LightGBM and CatBoost stood out with the highest Weighted F1 score and remarkable consistency in their performance. While XGBoost belongs to the same paradigm, its performance was anomalous in this experiment due to an observed technical incompatibility with the preprocessing pipeline, as noted in the table.

Based on these findings and considering its marginally superior performance and greater computational efficiency, LightGBM was chosen as the base algorithm for the hyperparameter optimization phase. This process, executed using GridSearchCV, allowed for fine-tuning its configuration, achieving a maximum F1 score of 0.8277 in cross-validation.

TABLE 4. MODEL BENCHMARKING RESULTS (DEFAULT CONFIGURATION)

Algorithm	F1-Weighted Score	Accuracy	Weighted Precision	Weighted Recall	ROC AUC (Multiclass)
LightGBM	0.8124	0.8133	0.8131	0.8133	0.9247
CatBoost	0.8123	0.8130	0.8128	0.8130	0.9243
Random Forest	0.8002	0.8008	0.8009	0.8008	0.9178
MLP (Neural Network)	0.7835	0.7854	0.7880	0.7854	0.9099
SVM	0.7789	0.7810	0.7857	0.7810	0.9049
Logistic Regression	0.7472	0.7498	0.7530	0.7498	0.8896
XGBoost (Anomalous)	0.0824	0.1930	0.3415	0.1930	0.5010

Source: Authors, 2025. *Note: The anomalously low performance of XGBoost is attributed to an observed incompatibility between the GPU implementation of the algorithm and the SMOTE up sampling pipeline used in the benchmark.*

The optimal hyperparameter settings found for the LightGBM model are detailed in Table 5.

TABLE 5. OPTIMAL HYPERPARAMETERS FOR THE FINAL LIGHTGBM MODEL

Hyperparameter	Optimal Value
learning_rate	0.05
n_estimators	300
num_leaves	31
reg_alpha	0.1
reg_lambda	0.1

Source: Authors, 2025.

This final model, called "Optimized LightGBM-SMOTE," which combines the best-performing algorithm with its optimal hyperparameter settings, is used for all performance evaluation and interpretability analysis in the following sections.

Evaluation of the Selected Final Model

The optimized LightGBM algorithm was evaluated on the test set, which consisted of 4,455 unused records. The evaluation was designed from a dual perspective: a technical analysis of the model in its standard configuration and a strategic analysis of its calibrated version.

The technical performance of the model, called the "Generalist Model," was measured using a decision threshold of 0.5. As detailed in Table 6, this model already demonstrates high-level performance, with an overall accuracy of 87% and an F1 score of 0.83 for the "Defected" class of interest.

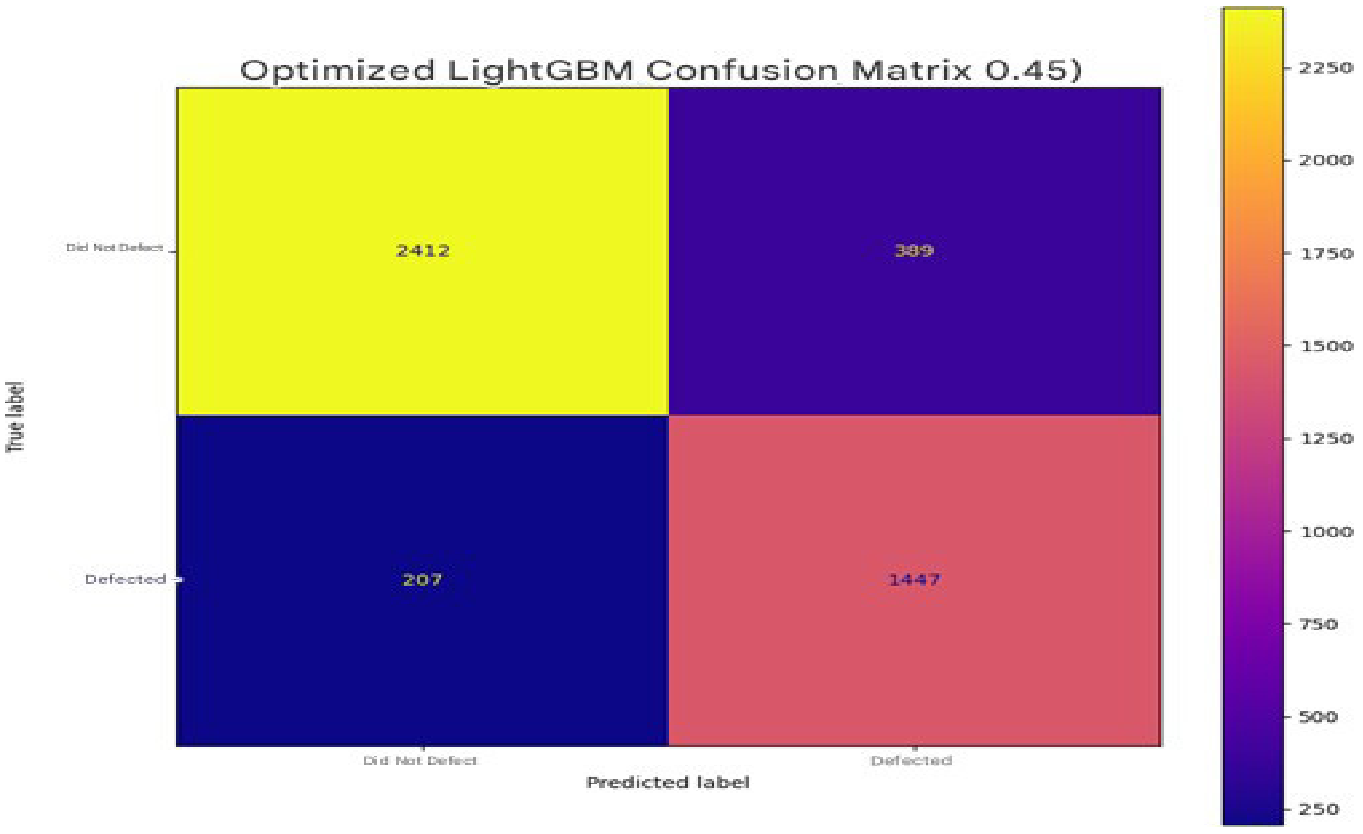
TABLE 6. CLASSIFICATION REPORT OF THE OPTIMIZED LIGHTGBM MODEL (THRESHOLD 0.5)

Class	Precision	Recall	F1-Score	Medium
Didn‘t Desert	0.92	0.87	0.89	2801
Deserted	0.80	0.86	0.83	1654
Accuracy			0.87	4455
Weighted Avg	0.87	0.87	0.87	4455

Source: Authors, 2025.

Subsequently, to more precisely align the model with the institutional objective of maximizing the detection of at-risk students, a “Strategic Model” was developed. This was obtained by calibrating the decision threshold to an optimal value of 0.45. The resulting Confusion Matrix is presented in Figure 3.

Fig 3. Confusion Matrix of the Strategic Model (LightGBM Calibrated to Threshold 0.45).



Source: Authors, 2025.

Analysis of the matrix reveals that the Strategic Model correctly identifies 1,447 of the 1,654 actual defectors, reducing false negatives to only 207 cases. The full performance profile of this calibrated model is presented in Table 7.

TABLE 7. STRATEGIC MODEL CLASSIFICATION REPORT (THRESHOLD 0.45)

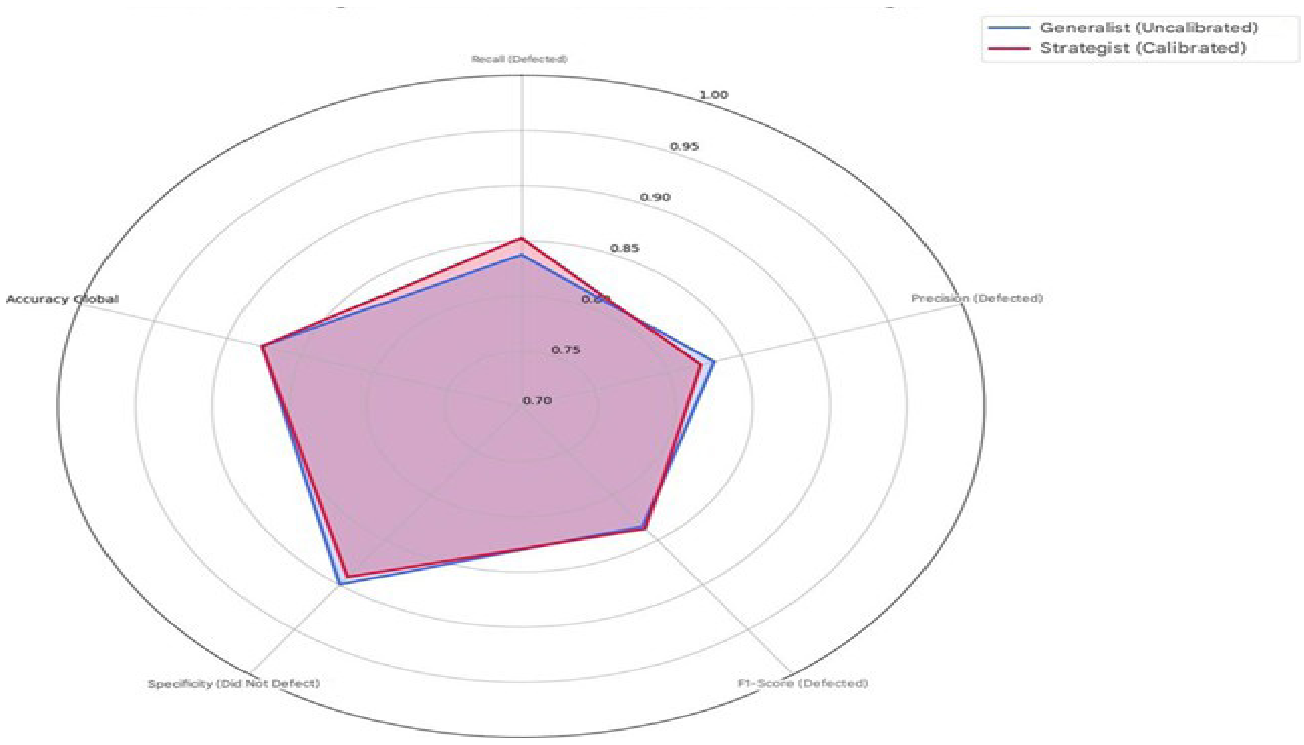
Class	Precision	Recall	F1-Score	Medium
Didn‘t Desert	0.92	0.86	0.89	2801
Deserted	0.79	0.87	0.83	1654
Accuracy			0.87	4455
Weighted Avg	0.87	0.87	0.87	4455

Source: Authors, 2025.

The comparison between the two models demonstrates the success of the calibration. The Strategic Model manages to increase the detection (recall) rate of the “Defected” class from 86% to 87%, while maintaining an F1 score of 0.83. This increase in sensitivity, achieved with a controlled trade-off in accuracy (which decreased from 80% to 79%), confirms the Strategic Model as the optimal and strategically superior solution for early intervention.

Finally, the result of this tactical adjustment is visualized in the Binary Strategic Radar (Figure 4), which directly compares the performance profiles of both models.

Fig 4. Strategic Radar comparing the performance profile of the standard model (Generalist) vs. the calibrated model (Strategist).



Source: Authors, 2025.

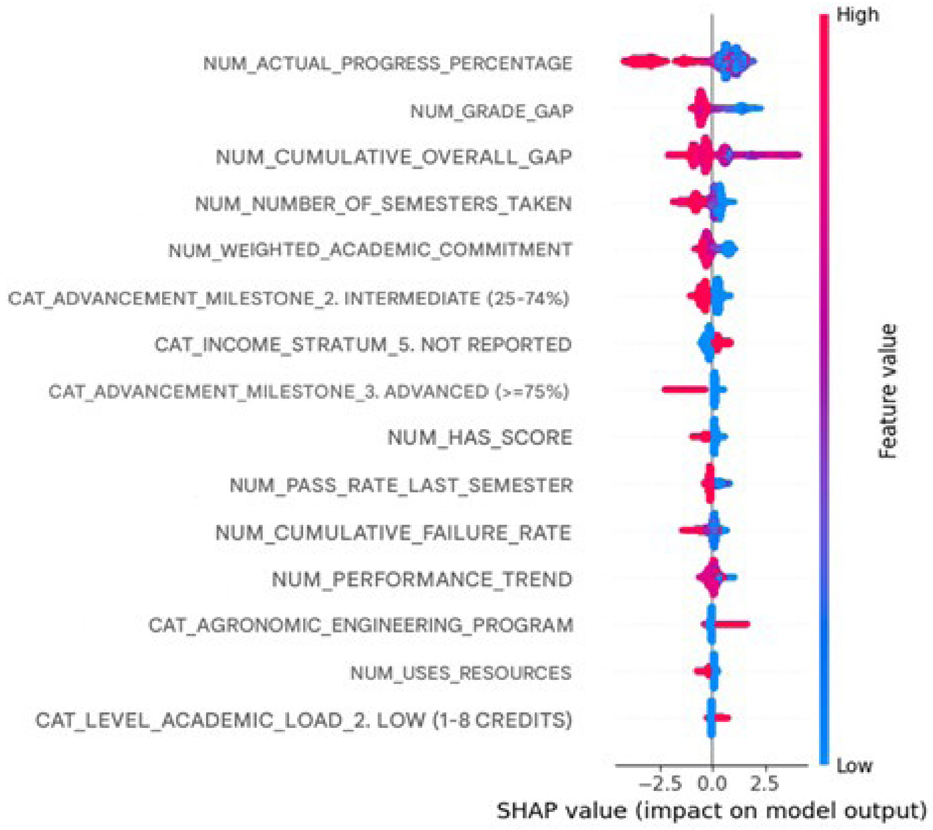
As the radar illustrates, the Strategist Model (in red) expands its reach along the Recall axis (the most critical mission metric) at the expense of a minimal and controlled contraction in Accuracy and Specificity, visually confirming its strategic superiority.

Model Interpretability Analysis (SHAP)

To overcome the “black box” nature of the LightGBM model and understand the factors that explain its predictions, the SHAP (Shapley Additive ExPlanations) methodology was applied. This approach allows us to quantify the contribution and direction of the impact of each variable on the model output.

Figure 5 presents a SHAP summary chart, which ranks the 15 most influential characteristics. In this chart, each point represents a student. The position on the x-axis indicates whether the value of a characteristic increases (SHAP value > 0) or decreases (SHAP value < 0) the predicted dropout risk. The color of the point indicates whether the characteristic value for that student is high (red) or low (blue).

Fig 5. Impact of characteristics on the prediction of the Dropout class. Each point represents a student.



Source: Authors, 2025.

For a clearer interpretation, Table 8 summarizes the direction of impact for the ten most important variables, as seen in the figure.

TABLE 8. INTERPRETATION OF THE DIRECTION OF IMPACT OF KEY VARIABLES

Range	Variable	Relationship with the Risk of Desertion
1	PERCENTAGE_ACTUAL_ PROGRESS	Negative. The further you advance in your career, the lower the risk of dropping out.
2	AVERAGE_GRADE_ GRADES	Negative. The higher the average in the last semester, the lower the risk.
3	CUMULATIVE_OVERALL_ AVERAGE	Negative. The higher the historical average, the lower the risk.
4	NUMBER_OF_ SEMESTERS_TAKEN	Negative. The greater the number of semesters, the lower the risk („sunk cost“ effect).
5	ACADEMIC_ COMMITMENT_ WEIGHTED	Negative. The higher the commitment (average x efficiency), the lower the risk.
6	PROGRESS_ MILESTONE_2. Intermediate	Negative. Being in the intermediate stage (vs. early) decreases the risk.
7	INCOME_STRATUM_5. Not Reported	Positive. Failure to report socioeconomic status increases the risk of dropping out.
8	PROGRESS_ MILESTONE_3. Advanced	Negative. Being in an advanced stage of your career drastically reduces the risk.
9	HAS_SCORE	Negative. Having a score (being an undergraduate) reduces the risk compared to those with a postgraduate degree (without a score).
10	LAST_SEM_APPROVAL_ RATE	Negative. The greater the efficiency in the last semester, the lower the risk.

Source: Authors (2025), based on SHAP analysis.

Analysis of the graph and table confirms that the model has learned logical and academically coherent patterns. It is conclusively revealed that the most dominant predictor is ACTUAL_ PROGRESS_PERCENTAGE, one of the engineering variables. Its operation is intuitive: a low progress percentage strongly pushes the prediction toward dropping out, while substantial progress acts as a protective factor. This same behavior is observed in the next most important predictors, such as GRADE_GPA and CUMULATIVE_OVERALL_GPA.

In addition to confirming the importance of performance, the SHAP results validate the feature engineering process. It is no coincidence that four of the variables created for this study figure prominently in the top 12 predictors. This confirms that the model effectively leveraged the contextual intelligence provided to it. Finally, the analysis shows that factors other than directly academic, such as the INCOME_STRATUM_5. Not Reported category, also play a measurable role, demonstrating that the model uses this context to refine its decisions.

DISCUSSION

El The final LightGBM model resulting from this study achieved remarkable performance. Its overall accuracy was 87%, but its true value lies in its ability to identify at-risk students. For the “Dropped Out” class, the model achieved a recall of 86%, a precision of 80%, and an F1-Score of 0.83. These metrics, obtained after a hyperparameter optimization that achieved an F1-Score of 0.8292 in cross-validation, demonstrate its robustness. Putting these results into context is crucial. How do they compare with other studies? Recent literature shows similar findings. For example, [64] reported an F1-Score of 0.840 with LightGBM, while [56] achieved 0.826 with the same algorithm, even outperforming Deep Neural Networks.

However, other machine learning models, including ensemble and deep learning models, have also reported high performance metrics, as is the case in the study by [5], which developed a predictive model with an accuracy of 90.3% in identifying at-risk students. Similarly, [63] proposed a student dropout prediction model with intuitionistic fuzzy sets and XGBoost (STOU2PM), achieving an accuracy of 91% and an area under the curve (AUC) of 0.913. Both results exceed the 87% accuracy of their own LightGBM model. On the other hand, in a study by [4], the Random Forest algorithm (80.56% accuracy) slightly outperformed XGBoost, which implies that XGBoost’s performance in that case was inferior to the presented LightGBM. It should be added that, [69] developed a classification model for student dropout with Gradient

Boosting Machine (GBM) using the H2O.ai framework, obtaining a Gini coefficient of 92.20% and an AUC of 96.10%, which suggests a highly efficient performance.

Meanwhile, a hybrid model called HLRNN (Hybrid Logistic Regression and Neural Network) proposed by [32] achieved 96% accuracy demonstrating its high predictive capacity. On the other hand, an ensemble model developed by [2], which combined Logistic Regression, Neural Networks and Decision Tree, correctly classified 89% of the students and accurately identified 98.1% of the dropouts (recall). In turn, other Artificial Neural Network (ANN) implementations by [49] achieved a classification accuracy of 90.3%, with the vulnerability index being the most relevant factor. Similarly, [53] reported that a 6-layer Multilayer Neural Network model obtained an accuracy of 98.97% in the training set; however, the best predictor identified by AutoAI of IBM Watson Studio was XGBoost with 87.1% accuracy, a value very close to that of the present study.

Regarding classical Machine Learning algorithms, [8] presented a Decision Tree model that obtained 92% precision and accuracy in determining the adaptability of students in online education. On the other hand, Logistic Regression, in the study by [49], showed a classification accuracy of 88.1%. Likewise, [6] concluded that Logistic Regression offered the best results to predict dropouts in their data set. Although Random Forest obtained an accuracy of 80.56% in the study by [4], [1] reported that a model based on the Analytic Hierarchical Process (AHP) had 64.6% predictability, while Logistic Regression reached 69.9%, values considerably lower than the current results. The differences in performance between the various models can be attributed to the heterogeneity of the data sets used, which include demographic, socioeconomic, academic and online behavior information. In addition, data preprocessing techniques, class imbalance management, and feature selection also significantly influence the results.

When reviewing the most relevant factors in university student dropout, the findings show that the REAL_PROGRESS_PERCENTAGE, the GPA and the CUMULATIVE_OVERALL_GPA are the most influential predictors, in agreement with [33], [36]. At the socioeconomic and demographic level, the absence of information on income is decisive, as indicated by [17], [20], [31]. Likewise, age, origin and employment status coincide with what was proposed by [42], [56]. Regarding personal and motivational factors, academic commitment and the use of institutional resources are related to what was stated by [18], [19], who highlighted resilience, self-efficacy and institutional interaction in the face of persistent demotivation or failure. Finally, the incorporation of Artificial Intelligence with explainable methods such as SHAP showed the relevance of four constructed variables, in line with [5], [50], [54], which confirms the multicausal nature of desertion and the value of AI to design focused retention strategies.

The study is limited to the analysis of a public higher education institution, which restricts the possibility of generalizing the results to private contexts. Furthermore, the data collected focused on academic, socioeconomic, and demographic variables, but, for reasons of personal information protection, did not include elements related to mental health problems or special educational needs (SEN – ISO 21001:2018). The exclusion of these components is due to the lack of systematized institutional records in these dimensions and to ethical restrictions on the handling of sensitive data, which reduces the explanatory capacity of the model in terms of the psychological and personal aspects that directly impact the continuation or abandonment of studies.

Future lines of research include the incorporation of databases from multiple universities with diverse academic profiles, the integration of qualitative variables related to student well-being and motivation, and the comparison of the LightGBM model with hybrid architectures or explainable neural networks. The use of text mining and sentiment analysis in unstructured sources, such as forums or surveys, is also proposed to broaden our understanding of the phenomenon. The development of intelligent systems integrated into academic platforms that generate real-time alerts, supporting decision-making and the design of more effective retention strategies, is recommended.

CONCLUSIONS

The study analyzed the factors associated with university dropout rates in Colombia based on a case study of 104,147 records from a Colombian university, where, after a rigorous screening process, three distinct categories of students were identified: continuing (44.1%), actual dropouts (37.1%), and graduates (18.8%). This finding confirmed that the percentage of actual progress, grade point average, and cumulative grade point average are the most decisive predictors of dropout, in interaction with socioeconomic variables such as income stratum and demographic variables such as age group. Based on this diagnosis, a predictive model was developed using artificial intelligence, specifically a LightGBM algorithm optimized with oversampling techniques (SMOTE), which achieved a weighted F1-Score of 0.83 for the "Dropout" class and an overall accuracy of 87%, demonstrating its ability to reliably identify students at risk. The evaluation of the model included technical, strategic, and interpretive phases: from standard metrics derived from the confusion matrix to the analysis of explainability using SHAP, with the aim of ensuring that the system did not operate as a black box, but rather that its prediction was based on clearly distinguishable patterns from the predictors included in the model.

Adjusting the prediction threshold to 0.45 was a decisive change that allowed us to move from an accuracy-focused approach to a model strategically oriented toward managing student dropout risk, increasing the sensitivity of the system and strengthening early detection. The calibration achieved a detection rate of 87% with 207 false negatives, generating a solid technical basis for more targeted institutional interventions aligned with retention and efficiency indicators.

The methodological pipeline incorporated data auditing, feature engineering, and hyperparameter adjustment, ensuring statistical stability and bias mitigation. This approach reduced leakage, strengthened interpretability, and improved the robustness of the model in real institutional settings. The results confirm that longitudinal academic trajectories have greater predictive power than static measurements, optimizing the anticipation of dropout events and the efficient allocation of resources.

The model not only predicts but also adapts to institutional objectives through strategic calibrations, integrating cost-benefit criteria and algorithmic governance. This capability positions it as an asset for early warning systems and educational decision-making processes, combining statistical accuracy, technical scalability, and operational relevance.

Limitations arising from the use of data from a single institution and the absence of psychometric variables open lines of research aimed at expanding the database and improving generalization. The use of advanced architectures such as stacking could optimize predictive stability, while an institutional pilot would allow the real impact of the model on dropout to be measured, strengthening strategic decision-making in higher education.

CRediT AUTHORSHIP CONTRIBUTION STATEMENT

A. Banquez-Maturana: Conceptualization, Methodology, Software, Validation, Formal analysis, Research, Resources, Data curation, Writing – Original draft, Writing – Review and editing, Visualization, Project management, Fund acquisition. **J. Rodríguez-Cerón:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Research, Resources, Data Curation, Writing – Original Draft, Writing – Review and Editing, Visualization, Project Management, Fundraising. **Á. Benavides-González:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Research, Resources, Data Curation, Writing – Original Draft, Writing – Review and Editing, Visualization, Project Management, Fundraising. **H. Felizzola Jimenez:** Methodology, Writing – Original Draft, Supervision.

FINANCING

This study was funded by the authors.

ACKNOWLEDGMENTS

The authors thank Universidad de La Salle for its support in the development of the research.

REFERENCES

[1] H. A. Silva, L. E. Quezada, A. M. Oddershede, P. I. Palominos, y C. O'Brien, «A Method for Estimating Students' Desertion in Educational Institutions Using the Analytic Hierarchy Process», *J. Coll. Stud. Retent. Res. Theory Pract.*, vol. 25, n.º 1, pp. 101-125, may 2023, doi: 10.1177/1521025120971227.

[2] A. M. Rabelo y L. E. Zárate, «A Model for Predicting Dropout of Higher Education Students», *Data Sci. Manag.*, jul. 2024, doi: 10.1016/j.dsm.2024.07.001.

[3] T. H. Nguyen, P. Le, T. T. T. Nguyen, y A. K. Su, «A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations», *CTU J Inn Sus Dev*, vol. 16, n.º Special issue: ISDS, pp. 98-106, oct. 2024, doi: 10.22144/ctujoisd.2024.327.

[4] L. G. R. Putra, D. D. Prasetya, y M. Mayadi, «Student Dropout Prediction Using Random Forest and XGBoost Method», *INTENSIF J. Ilm. Penelit. Dan Penerapan Teknol. Sist. Inf.*, vol. 9, n.º 1, pp. 147-157, feb. 2025, doi: 10.29407/intensif.v9i1.21191.

[5] A. Alhardi y S. Alan, *Predicting Student Dropout in Higher Education Using Machine Learning Techniques : A Predictive Model Using XGBoost Algorithm*. 2024.

[6] A. E. Aco Tito, B. Orlando Hancco Condori, y Y. Pérez Vera, «Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria», *RISTI Rev. Ibérica Sist. E Tecnol. Informação*, n.º 51, pp. 84-98, 2023.

[7] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén, y V. Cornejo-Aparicio, «Análisis del rendimiento académico de los estudiantes de Ingeniería de Sistemas, posibilidades de deserción y propuestas para su retención», *Ingeniare Rev. Chil. Ing.*, vol. 28, n.º 4, pp. 668-683, dic. 2020, doi: 10.4067/S0718-33052020000400668.

[8] L. E. A. Valencia, W. H. Condori, V. R. Q. Quicaña, y A. R. T. Coila, «Aplicación de árboles de decisión para la identificación de adaptabilidad de estudiantes en educación online», *Innov. Softw.*, vol. 4, n.º 2, pp. 166-181, sep. 2023, doi: 10.48168/innosoft.s12.a113.

[9] A. Behr, M. Giese, H. D. Teguim Kamdjou, y K. Theune, «Dropping out of university: a literature review», *Rev. Educ.*, vol. 8, n.º 2, pp. 614-652, jun. 2020, doi: 10.1002/rev3.3202.

[10] O. Lorenzo-Quiles, S. Galdón-López, y A. Lendínez-Turón, «Factors contributing to university dropout: a review», *Front. Educ.*, vol. 8, p. 1159864, mar. 2023, doi: 10.3389/feduc.2023.1159864.

[11] L. R. Pelima, Y. Sukmana, y Y. Rosmansyah, «Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review», *IEEE Access*, vol. 12, pp. 23451-23465, 2024, doi: 10.1109/ACCESS.2024.3361479.

[12] P. A. Willging y S. D. Johnson, «FACTORS THAT INFLUENCE STUDENTS' DECISION TO DROPOUT OF ONLINE COURSES», *Online Learn.*, vol. 13, n.º 3, feb. 2019, doi: 10.24059/olj.v13i3.1659.

[13] Nurmalitasari, Z. Awang Long, y M. Faizuddin Mohd Noor, «Factors Influencing Dropout Students in Higher Education», *Educ. Res. Int.*, vol. 2023, pp. 1-13, feb. 2023, doi: 10.1155/2023/7704142.

[14] D. A. Shafiq, M. Marjani, R. A. A. Habeeb, y D. Asirvatham, «Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review», *IEEE Access*, vol. 10, pp. 72480-72503, 2022, doi: 10.1109/ACCESS.2022.3188767.

[15] B. Albreiki, N. Zaki, y H. Alashwal, «A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques», *Educ. Sci.*, vol. 11, n.º 9, p. 552, sep. 2021, doi: 10.3390/educsci11090552.

[16] R. G. Venkatesan, D. Karmegam, y B. Mappillairaju, «Exploring statistical approaches for predicting student dropout in education: a systematic review and meta-analysis», *J. Comput. Soc. Sci.*, vol. 7, n.º 1, pp. 171-196, abr. 2024, doi: 10.1007/s42001-023-00231-w.

[17] V. R. D. C. Martinho, C. Nunes, y C. R. Minussi, «An Intelligent System for Prediction of School Dropout Risk Group in Higher Education Classroom Based on Artificial Neural Networks», en *2013 IEEE 25th International Conference on Tools with Artificial Intelligence*, Herndon, VA, USA: IEEE, nov. 2013, pp. 159-166. doi: 10.1109/ICTAI.2013.33.

[18] D.-M. Córdova-Esparza *et al.*, «Predicting and Preventing School Dropout with Business Intelligence: Insights from a Systematic Review», *Information*, vol. 16, n.º 4, p. 326, abr. 2025, doi: 10.3390/info16040326.

[19] J. G. C. Krüger, A. de S. Britto, y J. P. Barddal, «An explainable machine learning approach for student dropout prediction», *Expert Syst. Appl.*, vol. 233, p. 120933, dic. 2023, doi: 10.1016/j.eswa.2023.120933.

[20] Mst. R. Khatun, M. A. Mim, Md. M. Tasin, y Md. M. Hossain, «A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction», *PLOS One*, vol. 20, n.º 9, p. e0331917, sep. 2025, doi: 10.1371/journal.pone.0331917.

[21] M. Du, N. Liu, y X. Hu, «Techniques for interpretable machine learning», *Commun. ACM*, vol. 63, n.º 1, pp. 68-77, dic. 2019, doi: 10.1145/3359786.

[22] C. Molnar, G. Casalicchio, y B. Bischl, «Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges», en *ECML PKDD 2020 Workshops*, vol. 1323, I. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Guidotti, Ö. Özgöbek, R. P. Ribeiro, R. Gavaldà, J. Gama, L. Adilova, Y. Krishnamurthy, P. M. Ferreira, D. Malerba, I. Medeiros, M. Ceci, G. Manco, E. Masciari, Z. W. Ras, P. Christen, E. Ntoutsi, E. Schubert, A. Zimek, A. Monreale, P. Biecek, S. Rinzivillo, B. Kille, A. Lommatzsch, y J. A. Gulla, Eds., en Communications in Computer and Information Science, vol. 1323. , Cham: Springer International Publishing, 2020, pp. 417-431. doi: 10.1007/978-3-030-65965-3_28.

[23] S. M. Lundberg *et al.*, «From local explanations to global understanding with explainable AI for trees», *Nat. Mach. Intell.*, vol. 2, n.º 1, pp. 56-67, ene. 2020, doi: 10.1038/s42256-019-0138-9.

[24] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, y P. Cortadas, «An early warning system to identify and intervene online dropout learners», *Int. J. Educ. Technol. High. Educ.*, vol. 20, n.º 1, p. 3, ene. 2023, doi: 10.1186/s41239-022-00371-5.

[25] Opazo, Diego, S. Moreno, y E. Álvarez-Miranda, «Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities». Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: https://www.mdpi.com/2227-7390/9/20/2599

[26] S. Wang, F. Wang, Z. Zhu, J. Wang, T. Tran, y Z. Du, «Artificial intelligence in education: A systematic literature review», *Expert Syst. Appl.*, vol. 252, p. 124167, oct. 2024, doi: 10.1016/j.eswa.2024.124167.

[27] E. M.-C. Alamo, «Análisis de estrategias innovadoras para retención estudiantil con inteligencia artificial: una perspectiva multidisciplinaria», *epsir*, vol. 9, pp. 1-20, jul. 2024, doi: 10.31637/epsir-2024-440.

[28] N. Chiarino *et al.*, «Abandono y permanencia estudiantil en universidades de Latinoamérica y el Caribe: Una revisión sistemática mixta», *Act Inv En Educ*, vol. 24, n.º 2, pp. 1-37, may 2024, doi: 10.15517/aie.v24i2.57306.

[29] F. A. Al-azazi y M. Ghurab, «ANN-LSTM: A deep learning model for early student performance prediction in MOOC», *Heliyon*, vol. 9, n.º 4, p. e15382, abr. 2023, doi: 10.1016/j.heliyon.2023.e15382.

[30] P. G. Atangana Njock, S.-L. Shen, A. Zhou, y G. Modoni, «Artificial neural network optimized by differential evolution for predicting diameters of jet grouted columns», *J. Rock Mech. Geotech. Eng.*, vol. 13, n.º 6, pp. 1500-1512, dic. 2021, doi: 10.1016/j.jrmge.2021.05.009.

[31] A. F. Núñez-Naranjo, «Analysis of the determinant factors in university dropout: a case study of Ecuador», *Front Educ*, vol. 9, oct. 2024, doi: 10.3389/feduc.2024.1444534.

[32] S. Mustofa, Y. R. Emon, S. B. Mamun, S. A. Akhy, y M. T. Ahad, «A novel AI-driven model for student dropout risk analysis with explainable AI insights», *Comput. Educ. Artif. Intell.*, vol. 8, p. 100352, jun. 2025, doi: 10.1016/j.caeai.2024.100352.

[33] C. F. C. López y M. L. García, «Atención pedagógica a estudiantes con bajo rendimiento académico de primero de bachillerato general unificado», *Rev. Científica Tecnológica UPSE*, vol. 7, n.º 2, pp. 27-37, dic. 2020, doi: 10.26423/rctu.v7i2.506.

[34] G. Hägg y J. Gabrielsson, «A systematic literature review of the evolution of pedagogy in entrepreneurial education research», *Int. J. Entrep. Behav. Res.*, vol. 26, n.º 5, pp. 829-861, ene. 2020, doi: 10.1108/IJEBR-04-2018-0272.

[35] K. Ahmad, W. Iqbal, Ammar El-Hassan, y Junaid Qadir, «Data-Driven Artificial Intelligence in Education: A Comprehensive Review \textbar IEEE Journals & Magazine \textbar IEEE Xplore». Accedido: 23 de octubre de 2024. [En línea]. Disponible en: https://ieeexplore-ieee-org.hemeroteca.lasalle.edu.co/document/10247566/keywords#keywords

[36] Q. Fu, Z. Gao, J. Zhou, y Y. Zheng, «CLSA: A novel deep learning model for MOOC dropout prediction», *Comput. Electr. Eng.*, vol. 94, p. 107315, sep. 2021, doi: 10.1016/j.compeleceng.2021.107315.

[37] D. A. Lagunes-Ramírez, G. González-Serna, L. Rivera-Rivera, N. González-Franco, D. Mújica-Vargas, y M. Y. Hernández-Pérez, «Comportamiento de la mirada y análisis mediante aprendizaje automático de la depresión en la juventud: una revisión sistemática», *XIKUA Bol. Científico Esc. Super. Tlahuelilpan*, vol. 12, n.º 23, pp. 56-68, ene. 2024, doi: 10.29057/xikua.v12i23.11808.

[38] H. R. Paz, «College Dropout Factors: An Analysis with LightGBM and Shapley's Cooperative Game Theory», n.º arXiv:2311.06260. arXiv, 26 de septiembre de 2023. doi: 10.48550/arXiv.2311.06260.

[39] O. V. Quintana, «Combate a la corrupción desde el aula escolar (programa de apoyo a la educación)», *Rev. Científica Salud Desarro. Hum.*, vol. 5, n.º 2, pp. 1289-1304, ago. 2024, doi: 10.61368/r.s.d.h.v5i2.240.

[40] N. Simhadri y S. T. N. V. R., «Awareness among teaching on AI and ML applications based on fuzzy in education sector at USA \textbar Soft Computing». Accedido: 23 de octubre de 2024. [En línea]. Disponible en: https://link-springer-com.hemeroteca.lasalle.edu.co/article/10.1007/s00500-023-08329-z#additional-information

[41] Venegas, «Combate a la corrupción desde el aula escolar (programa de apoyo a la educación) \textbar Revista Científica de Salud y Desarrollo Humano». Accedido: 7 de septiembre de 2024. [En línea]. Disponible en: https://revistavitalia.org/index.php/vitalia/article/view/240

[42] A. Zanellati, S. P. Zingaro, y M. Gabbrielli, «Balancing Performance and Explainability in Academic Dropout Prediction», *IEEE Trans. Learn. Technol.*, vol. 17, pp. 2140-2153, 2024, doi: 10.1109/TLT.2024.3425959.

[43] A. J. Fernández-García, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, y F. Sánchez-Figueroa, «A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data», *IEEE Access*, vol. 9, pp. 133076-133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[44] M. K. Hossen y M. S. Uddin, «Attention monitoring of students during online classes using XGBoost classifier», *Comput. Educ. Artif. Intell.*, vol. 5, p. 100191, ene. 2023, doi: 10.1016/j.caeai.2023.100191.

[45] H. E. Caselli Gismondi y L. V. Urrelo Huiman, «Características para un modelo de predicción de la deserción académica universitaria. Caso Universidad Nacional de Santa», *LLamkasun Rev. Investig. Científica Tecnológica*, vol. 2, n.º 4, pp. 2-22, 2021.

[46] T. Susnjak, «Beyond Predictive Learning Analytics Modelling and onto Explainable Artificial Intelligence with Prescriptive Analytics and ChatGPT», *Int J Artif Intell Educ*, vol. 34, n.º 2, pp. 452-482, jun. 2024, doi: 10.1007/s40593-023-00336-3.

[47] A. Kuz y R. Morales, «Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México», *Educ. Knowl. Soc. EKS*, n.º 24, p. 13, 2023.

[48] M. C. Jiménez Mora, «Abandono y permanencia en educación superior: un análisis multinivel para Iberoamérica», Trabajo de grado - Maestría, Universidad Nacional de Colombia, 2021. Accedido: 20 de enero de 2025. [En línea]. Disponible en: https://repositorio.unal.edu.co/handle/unal/80950

[49] J. I. Vidal Chica, «Factores motivacionales, emocionales y socioeconómicos asociados al rendimiento académico y abandono en la Educación Superior: Estudio en dos Universidades del Ecuador», http://purl.org/dc/dcmitype/Text, Universitat d'Alacant / Universidad de Alicante, 2023. Accedido: 25 de agosto de 2024. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/tesis?codigo=312420

[50] W. Villegas-Ch, J. Govea, y S. Revelo-Tapia, «Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach», *Sustainability*, vol. 15, n.º 19, p. 14512, oct. 2023, doi: 10.3390/su151914512.

[51] V. Christou *et al.*, «Performance and early drop prediction for higher education students using machine learning», *Expert Syst. Appl.*, vol. 225, p. 120079, sep. 2023, doi: 10.1016/j.eswa.2023.120079.

[52] B. Alnasyan, M. Basheri, y M. Alassafi, «The power of Deep Learning techniques for predicting student performance in Virtual Learning Environments: A systematic literature review», *Comput. Educ. Artif. Intell.*, vol. 6, p. 100231, jun. 2024, doi: 10.1016/j.caeai.2024.100231.

[53] H. E. C. Gismondi y L. V. U. Huiman, «Multilayer Neural Networks for Predicting Academic Dropout at the National University of Santa - Peru», presentado en Proceedings - 7th International Symposium on Accreditation of Engineering and Computing Education, ICACIT 2021, 2021. doi: 10.1109/ICACIT53544.2021.9612507.

[54] D. Rivero-Albarrán, L. Guerra Torrealba, S. Arciniegas Aguirre, y O. Alexander, «Support System to Predict Student Dropout in Universities», en *Communication and Applied Technologies*, P. C. López-López, D. Barredo, Á. Torres-Toukoumidis, A. De-Santis, y Ó. Avilés, Eds., Singapore: Springer Nature, 2023, pp. 3-12. doi: 10.1007/978-981-19-6347-6_1.

[55] M. S. Asto-Lázaro y H. P. Bermejo-Terrones, «Systematic Review: Machine Learning in Academic Dropout Prediction», *RISTI - Rev. Iber. Sist. E Tecnol. Inf.*, vol. 2023, n.º E64, pp. 463-476, 2023.

[56] H. G. Kim, «Performance Comparison of Neural Network and Gradient Boosting Machine for Dropout Prediction of University Students», *J. Korea Soc. Comput. Inf.*, vol. 28, n.º 8, pp. 49-58, 2023, doi: 10.9708/jksci.2023.28.08.049.

[57] E. J. Brand C, V. M. Gabriel Ramirez, J. Diaz, y F. Moreira, «Toward Educational Sustainability: An AI System for Identifying and Preventing Student Dropout», *Rev. Iberoam. Tecnol. Aprendiz.*, vol. 19, pp. 100-110, 2024, doi: 10.1109/RITA.2024.3381850.

[58] H. S. Park y S. J. Yoo, «Early Dropout Prediction in Online Learning of University using Machine Learning», *JOIV Int. J. Inform. Vis.*, vol. 5, n.º 4, pp. 347-353, dic. 2021, doi: 10.30630/joiv.5.4.732.

[59] M. Chen y Z. Liu, «Predicting performance of students by optimizing tree components of random forest using genetic algorithm», *Heliyon*, vol. 10, n.º 12, p. e32570, jun. 2024, doi: 10.1016/j.heliyon.2024.e32570.

[60] D. C. Aráuz y J. J. Martínez, «Predicción del rendimiento académico en la UNADECA por medio de sistemas de clasificación», *Unaciencia Rev. Estud. E Investig.*, vol. 16, n.º 31, pp. 17-35, dic. 2023, doi: 10.35997/unaciencia.v16i31.738.

[61] T. Kakizaki y S. Oeda, «Student modeling considering learning behavior history with deep factorization machines», *Procedia Comput. Sci.*, vol. 225, pp. 2808-2815, ene. 2023, doi: 10.1016/j.procs.2023.10.273.

[62] K. Coussement, M. Phan, A. D. Caigny, D. F. Benoit, y A. Raes, «Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model», *Decis. Support Syst.*, vol. 135, p. 113325, 2020, doi: https://doi.org/10.1016/j.dss.2020.113325.

[63] W. Romsaiyud *et al.*, «Predictive Modeling of Student Dropout Using Intuitionistic Fuzzy Sets and XGBoost in Open University», en *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, en MLMI '24. New York, NY, USA: Association for Computing Machinery, dic. 2024, pp. 104-110. doi: 10.1145/3696271.3696288.

[64] C. H. Cho, Y. W. Yu, y H. G. Kim, «A Study on Dropout Prediction for University Students Using Machine Learning», *Appl. Sci.*, vol. 13, n.º 21, p. 12004, ene. 2023, doi: 10.3390/app132112004.

[65] Julio Elvis Valero Cajahuanca, «Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción», *RCS*, 2022, doi: 10.31876/rcs.v28i3.38480.

[66] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, y P. C. Nshimyumukiza, «Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization», *Comput. Educ. Artif. Intell.*, vol. 3, p. 100066, ene. 2022, doi: 10.1016/j.caeai.2022.100066.

[67] W. A. Bitencourt, D. M. Silva, y G. do Carmo Xavier, «May Artificial Intelligence support actions against school dropout?», *Ensaio*, vol. 30, n.º 116, pp. 669-694, 2022, doi: 10.1590/S0104-403620220003002854.

[68] R. Goran *et al.*, «Identifying and Understanding Student Dropouts Using Metaheuristic Optimized Classifiers and Explainable Artificial Intelligence Techniques», *IEEE Access*, vol. 12, pp. 122377-122400, 2024, doi: 10.1109/ACCESS.2024.3446653.

[69] H. Villarreal-Torres, J. Ángeles-Morales, W. Marín-Rodriguez, y J. Cano-Mejía, «Modelo de clasificación para la deserción estudiantil en las universidades públicas del Perú», *Rev. Cienc. Soc.*, vol. 30, n.º 1, pp. 452-469, 2024, doi: 10.31876/rcs.v30i1.41667.

[70] E.-Y. Seo, J. Yang, J.-E. Lee, y G. So, «Predictive modelling of student dropout risk: Practical insights from a South Korean distance university», *Heliyon*, vol. 10, n.º 11, p. e30960, jun. 2024, doi: 10.1016/j.heliyon.2024.e30960.

[71] N. B. Correa y M. A. L. Páez, «Regresión Logística Técnica de Machine Learning para predicciones académicas», *XIKUA Bol. Científico Esc. Super. Tlahuelilpan*, vol. 12, n.º Especial, pp. 71-80, jul. 2024, doi: 10.29057/xikua.v12iEspecial.12746.

[72] I. Elbouknify *et al.*, «Student At-Risk Identification and Classification Through Multitask Learning: A Case Study on the Moroccan Education System», *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 14830 LNAI, pp. 372-380, 2024, doi: 10.1007/978-3-031-64299-9_33.

[73] S. Kim, E. Choi, Y.-K. Jun, y S. Lee, «Student Dropout Prediction for University with High Precision and Recall», *Appl. Sci.*, vol. 13, n.º 10, p. 6275, ene. 2023, doi: 10.3390/app13106275.

[74] E. Melo, I. Silva, D. G. Costa, C. M. D. Viegas, y T. M. Barros, «On the Use of eXplainable Artificial Intelligence to Evaluate School Dropout», *Educ. Sci.*, vol. 12, n.º 12, p. 845, nov. 2022, doi: 10.3390/educsci12120845.

[75] H. Alonso-Misol Gerlache, P. Moreno Ger, y L. de la Fuente Valentín, «Towards the Grade's Prediction. A Study of Different Machine Learning Approaches to Predict Grades from Student Interaction Data», *IJIMAI*, vol. 7, n.º 4, pp. 196-204, 2022.

[76] T. M. Barros, P. A. Souza Neto, I. Silva, y L. A. Guedes, «Predictive Models for Imbalanced Data: A School Dropout Perspective», *Educ. Sci.*, vol. 9, n.º 4, p. 275, nov. 2019, doi: 10.3390/educsci9040275.

[77] D. Devi y S. Sophia, «GA-CNN: Analyzing student's cognitive skills with EEG data using a hybrid deep learning approach», *Biomed. Signal Process. Control*, vol. 90, p. 105888, abr. 2024, doi: 10.1016/j.bspc.2023.105888.

[78] J. K. Hoyos Osorio y G. Daza Santacoloma, «Predictive Model to Identify College Students with High Dropout Rates», *REDIE Rev. Electrónica Investig. Educ.*, n.º 25, pp. 1-10, 2023.

[79] M. Y. Amare y S. Šimonová, «Global Challenges of Students Dropout: A Prediction Model Development Using Machine Learning Algorithms on Higher Education Datasets», 2021, doi: 10.1051/shsconf/202112909001.

[80] C. Jin, «MOOC student dropout prediction model based on learning behavior features and parameter optimization», *Interact. Learn. Environ.*, vol. 31, n.º 2, pp. 714-732, feb. 2023, doi: 10.1080/10494820.2020.1802300.

[81] W. Castiblanco Vargas, L. R. Fonseca Gómez, y W. Pineda Ríos, «Detección de alertas tempranas para la prevención de la deserción estudiantil en una institución de educación superior a partir de un modelo de clasificación y su predicción por medio de técnicas de machine learning», *Conoc. Glob.*, vol. 6, n.º Extra 1, p. 43, 2021.

BIOGRAPHY OF AUTHORS

Ana Gabriela Banquez Maturana, Industrial Administrator and master's student in artificial intelligence, with academic and administrative experience in the Vice-Rector's Office of Quality Assurance at the University of Cartagena. She also works as a scientific researcher in the "Comprehensive Organizational Quality and Productivity" group (COL0048115), a scientific reviewer for high-impact journals classified in Q1–Q4 (ResearcherID HOA-8090-2023), and an advisor on editorial boards such as the Journal of Small Business and Enterprise Development (Q1), the Journal of Global Operations and Strategic Sourcing (T1), and the Journal of Health Organization and Management (Q2), among others. She is also a member of the Ibero-American Interdisciplinary Network of Researchers. https://orcid.org/0000-0002-8354-6396

Juan David Rodríguez Cerón, An electronic engineer, project management specialist, and soon-to-be master's student in artificial intelligence, he conducts his research in the field of artificial intelligence applied to education, particularly in predicting university dropout rates using machine learning approaches and longitudinal data analysis, integrating mathematical models, optimization techniques, multivariate analysis, and statistical control systems for decision-making. His experience extends to the management of rural electrification projects, teaching at different educational levels, and the incorporation of emerging technologies in academic settings, consolidating an interdisciplinary career that articulates technological innovation with scientific research to address highly relevant social and educational issues. https://orcid.org/0009-0004-9694-8476

Ángel Manuel Benavides González, A civil engineer, specialist in geographic information systems, and soon-to-be master's degree in artificial intelligence, he focuses on the use of artificial intelligence in education, developing a model for predicting university dropout rates using machine learning approaches and longitudinal data analysis, integrating mathematical models, optimization techniques, multivariate analysis, and statistical control systems for decision-making. His experience spans geomatics, remote sensing, and spatial analysis. https://orcid.org/0000-0002-2890-500X

Heriberto Alexander Felizzola Jimenez, Industrial Engineer, Master's in Industrial Engineering, and PhD candidate in Engineering from the Universidad de los Andes. He is currently an Assistant Professor in the Department of Innovation, Automation, and Productivity at the Universidad de La Salle. His career combines more than 15 years of experience in undergraduate and graduate university teaching with solid research and consulting work in data science, operations research, and process improvement through Lean Six Sigma. He has led research projects focused on open data analysis in public procurement, with an emphasis on risk prediction and promoting transparency in Colombia, as well as on process optimization in the productive, service, and healthcare sectors. His academic output includes publications in high-impact indexed journals (Q1–Q4) and participation in international conferences. He has also supported training and knowledge transfer processes in public and private companies and institutions in Colombia, applying analytical and management methodologies for data-driven decision-making. His work is characterized by the integration of advanced analytics tools and industrial engineering approaches to generate innovative solutions to complex problems in operations, education, and public policy. https://orcid.org/0000-0003-3149-8182

TABLE A1. ABBREVIATIONS AND FULL NAMES OF MACHINE LEARNING TECHNIQUES

Abbreviation	Name of Techniques	Number of items
AdaB	AdaBoost (AGbSCHO)	2
AML	Attention-based Multi-layer – Long Short-Term Memory	1
ANN	Artificial Neural Networks	7
ANN-LSTM	Artificial Neural Network – Long Short-Term Memory	1
BAG	Bagging (Bootstrap Aggregation)	1
BFGS-BP	Broyden–Fletcher–Goldfarb–Shanno Backpropagation	1
BN	Bayesian Network	1
CART-D	Classification And Regression Tree Discretized	1
CART-N	Classification And Regression TreeNumeric	1
CAT	CatBoost	1
DeepFM	Deep Factorization Machine	1
DFFNN	Deep Feed Forward Neural Network	1
DKT	Deep Knowledge Tracing	2
DKVMN	Dynamic Key-Value Memory Network	1
DNN	Deep Neural Network	3
DT	Decision Tree	11
ENS	Ensemble (GB + RF + SVM)	5
FNN	Feed Forward Neural Network	1
FTT	Feature Tokenizer Transformer	1
GB	Gradient Boosting	4
GBT	Gradient Boosted Trees	1
GCA-NN	Graph-based Conventional Attention Neural Network	1
HLRNN	Hierarchical Layer Recurrent Neural Network	1
IF	Isolation Forest	1
K-M	K-means	1
KNN	K-Nearest Neighbors	2
LGBM	Light Gradient Boosting Machine	6
LIME	Local Interpretable Model-agnostic Explanations	2
LM-BP	Levenberg–Marquardt Backpropagation	1
LR	Logistic Regression	10
MLP	Multilayer Perceptron	4
MLR	Multilevel Logistic Regression	1
NB	Naïve Bayes	4
NNC	Neural Network Construction	1
RBFNN	Radial Basis Function Neural Network	1
RF	Random Forest	17
RF+GA	Random Forest + Genetic Algorithm	1
RF-ent	Random Forest (‘entropy’)	1
RLN	Linear Regression	1
RNN	Recurrent Neural Network	1
SAKT	Self-Attentive Knowledge Tracing	1
SDP	SDP System (Hybrid: XGBoost + CatBoost)	1
SGD	Stochastic Gradient Descent	1
SHAP	SHapley Additive exPlanations	4
SVM	Support Vector Machine	6
SVRQ	Support Vector Regression with Improved Quantum Particle Swarm Optimization	1
XAI	eXplainable Artificial Intelligence	1
XGB	Extreme Gradient Boosting (XGBoost)	9
XGB+IFS	XGBoost + Intuitionistic Fuzzy Sets	1

Source: Authors, 2025.