IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 628∼641
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp628-641

Predicting university student dropouts in Latin America using machine learning

Laberiano Andrade-Arenas[1], Inoc Rubio Paucar[2], Margarita Giraldo Retuerto[1], Cesar Yactayo-Arias[3]
[1]Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú
[2]Facultad de Ingeniería y Negocios,Universidad Privada Norbert Wiener, Lima, Perú
[3]Departamento de Estudios Generales, Universidad Continental, Lima, Perú

Article Info

Article history:

Received Aug 14, 2025
Revised Dec 29, 2025
Accepted Jan 22, 2026

Keywords:

Decision making
Machine learning
Predictive model
Random forest
Student dropout

ABSTRACT

In the university context, student dropout has become one of the most recurring problems, both in the short and long term. The objective of this research was to develop a predictive model using the random forest (RF) algorithm to identify patterns associated with university dropout. To achieve this, the knowledge discovery in databases (KDD) methodology was applied, which encompasses the stages of selection, preprocessing, transformation, data mining, and interpretation of results. The RF model demonstrated superior performance compared to other evaluated models, achieving an accuracy of 87%, a precision of 86%, a recall of 85%, an F1-score of 85%, and an receiver operating characteristic (ROC) area under the curve (AUC) of 0.91, highlighting its high predictive capability compared to other techniques analyzed. Therefore, the application of the proposed model is recommended in various university institutions in order to identify potential dropout cases at an early stage.

Corresponding Author:

Laberiano Andrade-Arenas
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades
Lima, Perú
Email: landrade@uch.edu.pe

1. INTRODUCTION

In the current context, university student dropout has become one of the most pressing global issues, with both social and economic implications. High dropout rates limit students' professional development, reduce the efficiency of higher education institutions, and directly impact the growth and competitiveness of countries. This situation not only represents a loss of talent and resources but also undermines efforts to ensure quality education [1], [2]. It is essential to take action in response to this situation, as student dropout has become a common occurrence in universities, driven by multiple factors.

Despite institutional efforts to improve educational quality, student dropout in higher education remains a persistent and multifactorial challenge. The causes of dropout are diverse and include academic, personal, economic, and contextual factors, which make timely identification difficult through traditional methods. This complexity prevents many universities from anticipating dropout risk and implementing effective interventions in a timely manner [3], [4]. Moreover, the limited availability of resources to carry out individualized student monitoring further complicates the implementation of appropriate preventive strategies. This situation not only affects institutional performance and educational planning, but also represents a significant loss of human talent, public investment, and personal and professional development opportunities for students. In addition, the emotional and motivational impact of dropping out can affect students' self-esteem,

creating a negative effect on their family and social environment [5], [6]. Therefore, it is urgent to strengthen academic support policies, guidance, and comprehensive assistance that can help address this issue from a more human and inclusive perspective. It is considered essential to approach student dropout with greater attention, as it represents a significant loss not only for students, but also for institutions and society as a whole.

This research is justified by the urgent need to reduce dropout rates in higher education—a problem that negatively impacts students, institutions, and national development. The multifactorial causes of dropout make early detection difficult through traditional methods, limiting the implementation of effective preventive strategies [7], [8]. In this context, it becomes essential to have tools that allow for the analysis of large volumes of academic data and the generation of accurate predictions regarding dropout risk. Machine learning emerges as an innovative and effective alternative for this purpose, as it enables the construction of predictive models capable of identifying risk patterns based on available data. This research will contribute to the development of decision-support systems in universities, facilitating timely and personalized interventions to improve student retention and promote academic success [9], [10]. Anticipating student dropout is essential, as timely intervention not only enhances academic performance but also provides greater opportunities for students' personal and professional development. The objective of this research is to develop a predictive model based on the random forest (RF) algorithm to identify patterns of student dropout, with the aim of optimizing strategic decision-making in the university context.

2. LITERATURE REVIEW

This section presents a thorough review of various studies related to the topic addressed. With the purpose of providing a broad and well-founded perspective on the subject of study. Additionally, the theoretical frameworks consulted support the selection and interpretation of the variables considered in the analysis.

2.1. Related works

This research proposes a machine learning-based approach for evaluating teaching performance. To address this issue, several classification algorithms were implemented using the Python programming language, including k-nearest neighbors (KNN), extra trees, light gradient boosting machine (LightGBM), CatBoost classifier, among others. The results showed that the proposed model achieved a 2% higher accuracy compared to the other evaluated algorithms, highlighting its effectiveness in the educational context. In a complementary area, a student dropout prediction system was developed using machine learning algorithms, based on a longitudinal dataset collected from university students. The results indicated that the risk of dropout is primarily associated with factors such as academic department, gender, and socioeconomic group [11], [12]. Another relevant aspect addressed by Niyogisubizo *et al.* [13] was the proposal of a hybrid dropout prediction model, which combines the RF, extreme gradient boosting (XGBoost), gradient boosting (GB), and feedforward neural networks (FNN) algorithms. The model's performance was evaluated using the area under the curve (AUC), showing promising results in identifying factors related to school dropout. The analysis highlighted the impact of uncontrolled behaviors as a key variable in dropout risk. On the other hand, Vives *et al.* [14] emphasizes the effectiveness of long short-term memory (LSTM) networks in predicting academic performance. Through comparisons between different models based on metrics such as accuracy, precision, recall, and F1-score, the superiority of the LSTM - generative adversarial networks (GAN) model was confirmed, achieving an accuracy of 98.3% in week 8, followed by the deep neural networks (DNN) - GAN model with 98.1%.

In the context of predicting dropout in postgraduate programs, classification models such as logistic regression, RF, and neural networks were developed and optimized using resampling techniques to address class imbalance (synthetic minority over-sampling technique (SMOTE), SMOTE - support vector machine (SVM), adaptive synthetic (ADASYN)), as well as through hyperparameter tuning. The best-performing model was the neural network combined with SMOTE-SVM, achieving a recall value of 0.75, followed by logistic regression with 0.67 and RF with 0.60—the latter also demonstrating strong generalization ability with an optimal decision threshold of 0.427. Complementarily, another study focused on student dropout implemented a predictive model based on LightGBM, which achieved outstanding performance with an F1-score of 0.840, surpassing the results of previous studies that addressed the class imbalance issue. The model's effectiveness was enhanced through the application of oversampling techniques such as SMOTE, ADASYN, and Borderline-SMOTE, which helped improve class distribution and optimize the system's predictive capacity, as noted in [15], [16].

Another application oriented toward virtual learning environments adopted a hybrid approach using machine learning algorithms—specifically RF and XGBoost—to classify students at risk of dropping out. The

model achieved outstanding results, with an accuracy of 93%, a precision of 91.52%, a recall of 96.42%, and an F1-score of 93.91%, demonstrating its high effectiveness in the early detection of academic dropout. A further relevant contribution related to student dropout involved the development of a university dropout prediction system. For this purpose, a software prediction program was created based on machine learning models to identify the correlation between variables and student dropout. The models were evaluated for accuracy, with artificial neural networks of the perceptron type achieving the highest accuracy at 98.1% [17], [18]. Recent studies have developed a university dropout prediction system that significantly improved accuracy (0.963) and recall rate (0.766) by using dimensionality reduction techniques with principal component analysis (PCA) and clustering through K-means. The model outperformed the best previous approach by 0.093 in accuracy and achieved an F1-score of 0.808, surpassing the GB method. Additionally, it identified four main causes of dropout: employment, non-registration, personal problems, and admission to another university—the latter being the most accurately predicted (0.672). In a separate study, a classification model was implemented using machine learning techniques to anticipate student dropout with high levels of accuracy. The proposal followed a technological methodology with a propositional focus, incremental innovation, and synchronous scope. Data collection was conducted through a 20-question survey administered to 237 postgraduate students enrolled in education master's programs. The model, based on gradient boosting machine (GBM), yielded outstanding results: a Gini coefficient of 92.20%, an AUC of 96.10%, and a LogLoss of 24.24%. These results enabled the effective identification of key factors behind student dropout and provided a strategic tool for educational management [19], [20].

In a relevant alternative approach, the research in [21], [22] applied data mining techniques using academic grades as key predictive variables, combined with various machine learning algorithms aimed at modeling university dropout. The results demonstrated strong model performance, achieving an F1-score of 81% on the final test set. These findings suggest that students' academic performance is a representative indicator of their living conditions and, therefore, allows for the early detection of potential dropout cases in higher education. This supports the idea that academic success is influenced by multiple factors, including class imbalance, which justifies the use of supervised machine learning algorithms such as decision trees (DT) and SVM. However, boosting algorithms—especially LightGBM and CatBoost optimized with Optuna—showed superior performance compared to traditional classifiers, establishing themselves as more effective approaches for academic prediction, as highlighted by the aforementioned author. In another instance, when analyzing dropout risk among undergraduate students, unsupervised clustering algorithms were applied alongside RF and probability threshold adjustment. The traditional model yielded a low accuracy of 13.2% in predicting dropout, compared to 99.4% in retention. However, after adjusting the threshold, the accuracy in detecting dropout exceeded 50%, while maintaining overall and retention rates above 70% [23], [24].

This research addresses dropout in massive open online courses (MOOCs), proposing the use of the RF algorithm to predict this phenomenon. The model demonstrated strong performance, achieving an accuracy of 87.5%, an AUC of 94.5%, a precision of 88%, a recall of 87.5%, and an F1-score of 87.5%, highlighting its effectiveness in the early detection of university students at risk of dropping out. In addition, risk factors associated with dropout in university programs were identified by applying various machine learning algorithms, among which RF exhibited the most notable performance. The highest level of predictive accuracy was reached at the end of the first semester, once sufficient academic information about the students had been collected. At this stage, the model produced performance indicators that were comparable to those reported in previous research on early identification of dropout risk and low academic achievement [25], [26]. Several studies highlight the relevance of applying machine learning techniques in this context, particularly models such as GB, RF, and SVM, which have shown promising results for supporting institutional decision-making and for designing preventive strategies in university settings [27].

2.2. Student dropout

Student dropout in universities is defined as the student's decision to interrupt their studies for various context-related reasons, whether the interruption is temporary or permanent. Dropout represents a critical issue for universities, as it impacts the efficiency of the educational system, the allocation of resources, and the development of qualified human capital [28], [29]. This phenomenon arises from multiple causes, as outlined in Table 1, which seek to address this challenge. In this regard, it is advisable to closely monitor university students' academic performance, as it can significantly influence their long-term professional success or failure.

Table 1. Main causes of university student dropout

Category	Specific cause	Example	Impact type
Academic	Low performance	Continuous failure of courses	Academic
Economic	Lack of resources	Can't afford tuition or transportation	Economic
Vocational	Demotivation	Insecurity about career choice	Emotional/Vocational
Familiar	Family problems	Conflicts or responsibilities at home	Psychological
Institutional	Lack of mentoring	Poor academic support	Institutional
Social	Discrimination or exclusion	By gender, race or social class	Social/Cultural
Health	Medical or psychological problems	Anxiety, depression, chronic illnesses	Staff
Labor	Need to work	Drop out of school to work full-time	Economic/Labor

2.3. Random forest

The RF algorithm is a supervised machine learning method based on ensemble techniques, which involves building multiple independent DTs and combining their predictions to obtain more robust, accurate, and generalizable results. This model uses the bagging method, where each tree is trained on a random sample of the dataset, and at each split in the tree, a random subset of features is considered, which helps reduce the correlation between trees. For classification tasks, the final result is determined by majority voting, while for regression tasks, it is calculated by averaging the predictions. This approach improves performance by reducing overfitting and efficiently handles large volumes of data. However, its main drawback is its lower interpretability compared to a single DT [30], [31]. This type of algorithm can be applied in various contexts—such as medicine, education, and finance depending on the domain in which it is used.

3. METHODOLOGY

The knowledge discovery in databases (KDD) methodology is a comprehensive and systematic process aimed at transforming large volumes of raw data into useful, novel, understandable, and relevant knowledge for decision-making. This process includes several interrelated stages: the selection of relevant data, cleaning and preprocessing to remove inconsistencies or outliers, transformation into suitable formats, application of data mining techniques to extract meaningful patterns, and finally, the evaluation, interpretation, and presentation of the discovered knowledge in a way that can be understood and used by organizations [32], [33]. In this study, the process is applied to an institutional dataset composed of academic, socio-economic, and demographic student records, involving approximately 510 students. Before modeling, the data underwent a cleaning procedure, treatment of missing values, detection of outliers, and normalization to ensure analytical reliability. Figure 1 presents the phases of the KDD process, illustrating the data flow toward obtaining relevant results that support informed decision-making. Meanwhile, Figure 2 shows the architecture implemented for student data analysis. The workflow starts with the ingestion of datasets in formats such as .CSV, .XLSX, .TXT, and .JSON, processed using Python. The architecture integrates libraries such as Scikit-learn, XGBoost, NumPy, and Pandas, applying preprocessing steps including cleaning, standardization, and transformation. The RF model was subsequently implemented, allocating 80% of the dataset for training and the remaining 20% for validation. Finally, the model is evaluated through metrics such as accuracy, confusion matrix, F1-score, precision, recall, and ROC–AUC curves, aiming to obtain meaningful results that contribute to decision-making in educational contexts.

3.1. Selection

This section presents a thorough search focused on selecting the most appropriate dataset for the development of the machine learning project. The selection was based on the research objective, prioritizing data relevance, quality, and availability. To achieve this, several specialized platforms for public dataset distribution were explored, with Kaggle standing out as a leading platform due to its robustness and wide variety of datasets from different fields of knowledge. Kaggle is a reliable and up-to-date source, supported by an active scientific community that shares high-quality data along with detailed technical descriptions [34]. This feature allowed for the selection of a dataset aligned with the project's goals, ensuring a solid foundation for subsequent analysis, preprocessing, and modeling using machine learning techniques such as RF. It is important to note that Kaggle offers datasets across various domains and hosts competitions and publications centered on machine learning.

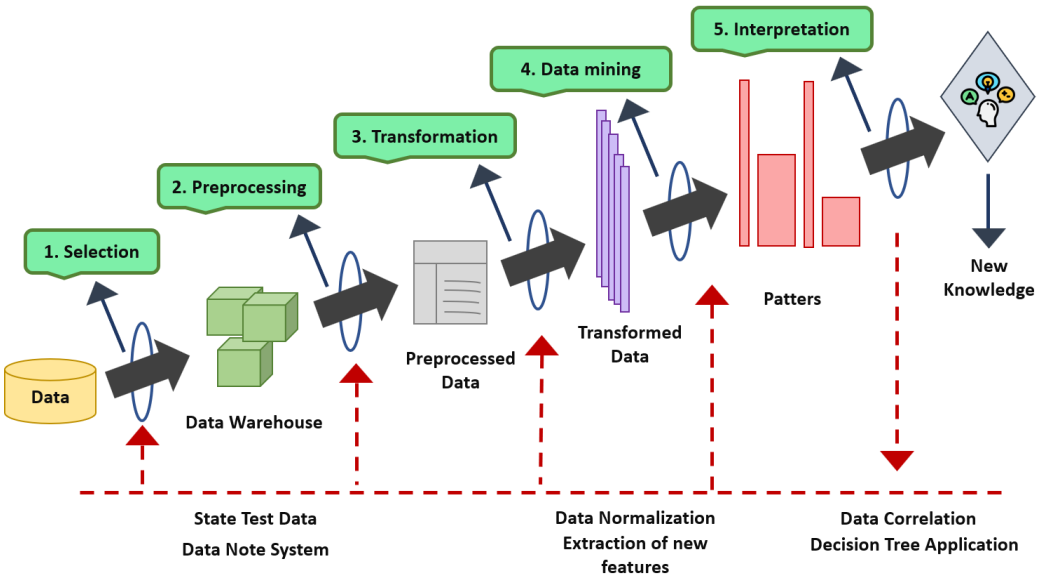


Figure 1. KDD methodology

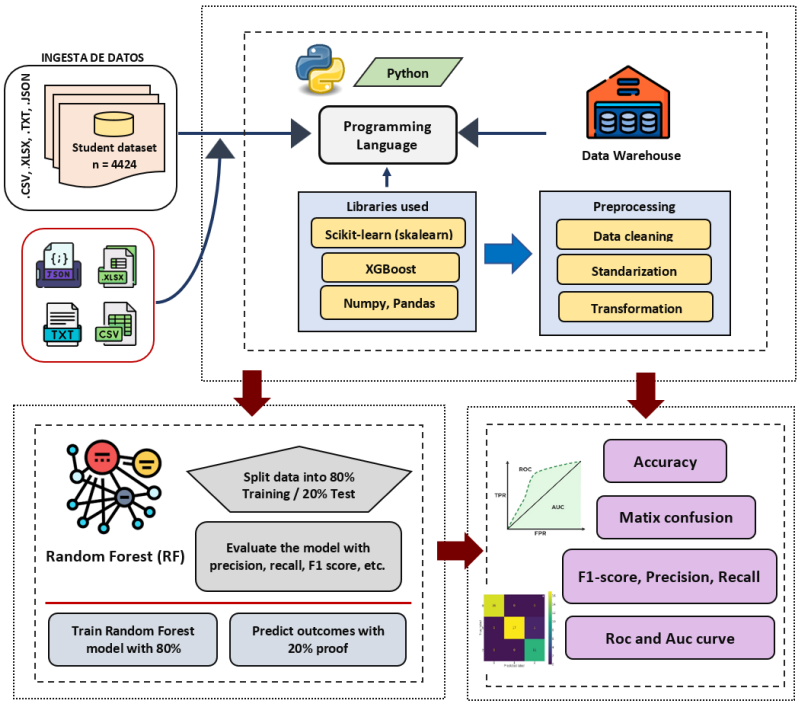


Figure 2. Machine learning architecture

3.2. Preprocessing and transformation

This section presents the preprocessing and transformation of the data. Tables 2 to 5 shows the results of the exploratory data analysis and the initial stages of variable preparation for the predictive model. Table 2 displays the analysis of missing values, where low percentages of missing data are observed 3.75% in mother's occupation and 4.01% in father's occupation. Variables such as debtor, tuition payment, and unemployment rate do not contain any missing values. Table 3 details the distribution of participants by marital status, with "single" being the most common category, followed by married, contributing to the sociodemographic profile of the

study. Table 4 presents the correlation between economic indicators, revealing negative relationships between gross domestic product (GDP) and both the unemployment rate (-0.40) and the inflation rate (-0.55), suggesting a link between economic growth and improved social conditions. Lastly, Table 5 shows the discretization of GDP into three levels (low, medium, and high), with the medium level being the most frequent. This facilitates its integration into classification models such as RF. These tables help to understand how the data is structured and what transformations are applied prior to modeling.

Table 2. Exploratory data analysis and preprocessing results, missing data analysis

Variable	Missing (%)
Marital status	2.14
Day/Night attendance	0.00
Mother's occupation	3.75
Father's occupation	4.01
Debtor	0.00
Tuition payment	0.00
International student	0.27
Unemployment rate	0.00
Inflation rate	0.00
GDP	0.00

Table 3. Exploratory data analysis and preprocessing results, marital status distribution

Category	Frequency
Single	210
Married	125
Divorced	30
Widowed	9
Total	374

Table 4. Exploratory data analysis and preprocessing results, correlation between economic indicators

	Unemployment rate	Inflation rate	GDP
Unemployment rate	1.00	0.65	-0.40
Inflation rate		1.00	-0.55
GDP			1.00

Table 5. Exploratory data analysis and preprocessing results, discretization of GDP values

Category	GDP range	Frequency
Low GDP	$<$ 10000	80
Medium GDP	10000–30000	200
High GDP	$>$ 30000	94
Total		374

Table 6 presents the descriptive statistics of the numerically encoded quantitative variables in the dataset. These statistics provide an overview of the behavior of the sociodemographic and economic variables considered in the study. The variable marital status has a mean value of 1.78, indicating that most participants fall between the categories of single and married. Similarly, the mean value for attendance (day or night) is 1.25, suggesting a higher proportion of students attending daytime classes. Parental occupation variables show average values close to 1.5, reflecting an intermediate distribution among employed, unemployed, or "other" categories. Regarding binary variables such as debtor, tuition payment, and international student, the low mean values indicate that most individuals are not in debt, are up to date with tuition payments, and are not international students, respectively. On the other hand, economic indicators reveal an average unemployment rate of 6.20%, an inflation rate of 2.45%, and a GDP average of 21,500.75 monetary units. These values help to understand the economic context in which the participants are situated and provide a solid foundation for further analysis. Overall, the statistical information of these variables facilitates data preparation and attribute selection for the construction of predictive models.

Table 6. Descriptive statistics of selected variables

Variable	Count	Mean	Std. Dev.	Min	Q1 (25%)	Q2 (Median)	Q3 (75%)	Max
Marital status	366.00	1.78	0.89	1.00	1.00	2.00	2.00	4.00
Day/Night attendance	374.00	1.25	0.43	1.00	1.00	1.00	1.00	2.00
Mother's occupation	360.00	1.65	0.75	1.00	1.00	2.00	2.00	3.00
Father's occupation	359.00	1.52	0.70	1.00	1.00	1.00	2.00	3.00
Debtor	374.00	0.25	0.43	0.00	0.00	0.00	1.00	1.00
Tuition payment	374.00	0.20	0.40	0.00	0.00	0.00	0.00	1.00
International student	373.00	0.09	0.29	0.00	0.00	0.00	0.00	1.00
Unemployment rate	374.00	6.20	1.45	3.20	5.10	6.00	7.30	9.80
Inflation rate	374.00	2.45	0.65	1.00	2.00	2.40	2.90	4.10
GDP	374.00	21500.75	7800.42	8000.00	16000.00	21000.00	27000.00	42000.00

3.3. Data mining

For this procedure, Figure 3 illustrates the architecture underlying the decision-making process within the DT framework, based on the dataset used. Meanwhile, Figure 4 displays the results that allow the evaluation of the RF model's performance in predicting student dropout. Figure 4(a) shows the ROC curve, showing the relationship between the true positive rate and the false positive rate, with a high AUC indicating strong discrimination between students who drop out and those who do not. Figure 4(b) presents the precision–recall curve, showing the balance between precision and recall, including the AUC value and the optimal threshold, which is especially useful in scenarios involving class imbalance. Figure 4(c) illustrates the relationship between sensitivity and specificity across different classification thresholds. As the threshold increases, sensitivity decreases while specificity increases. The intersection point of the two curves at approximately a threshold of 0.4 suggests a possible balance between these metrics. The legend includes the formulas, and the sidebar indicates the threshold values.

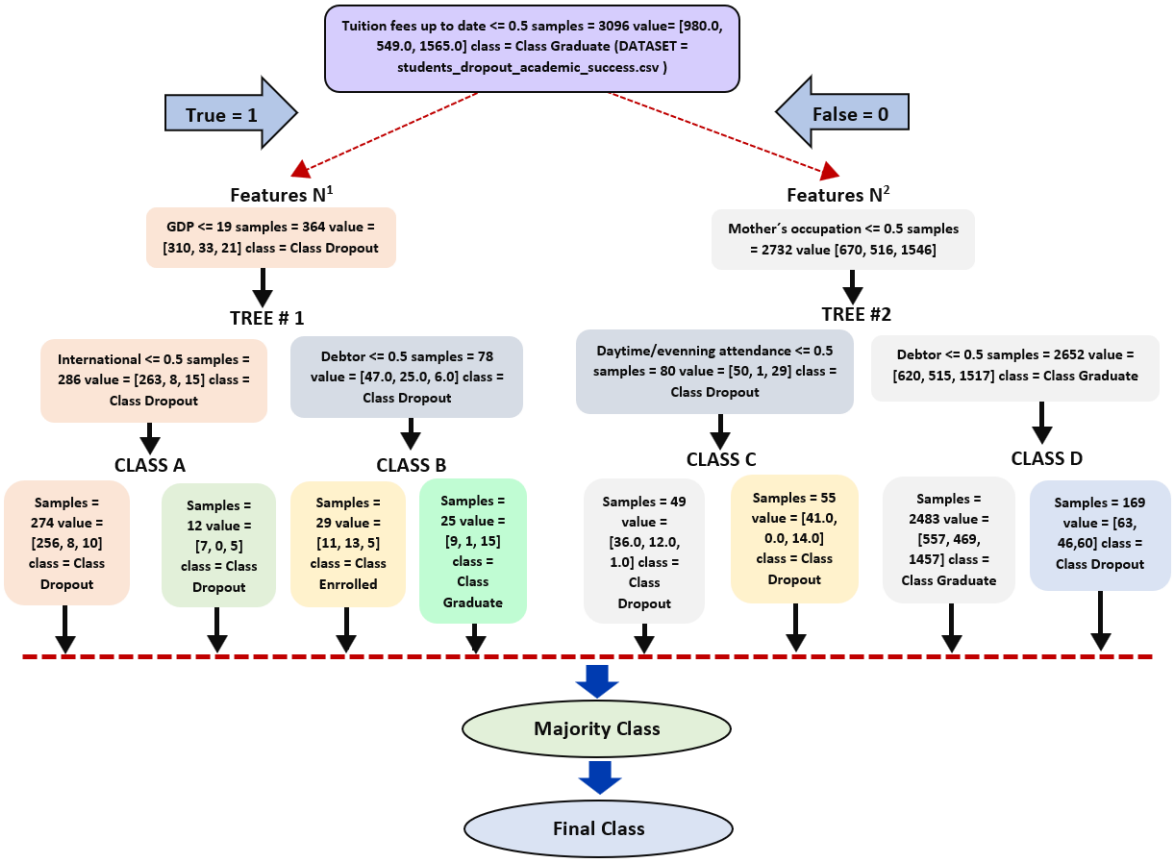
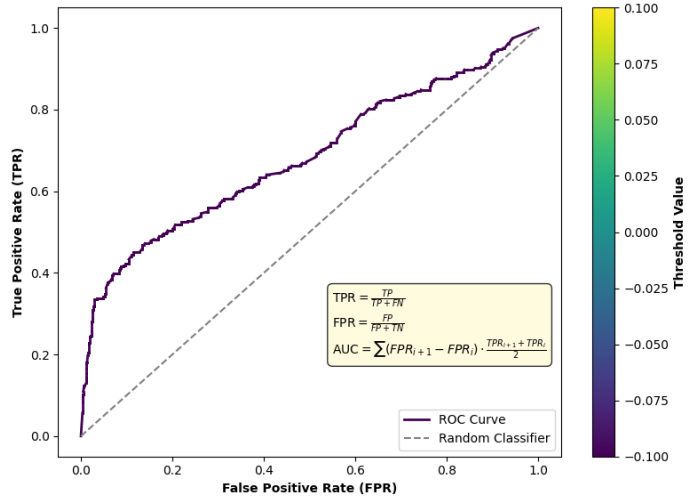
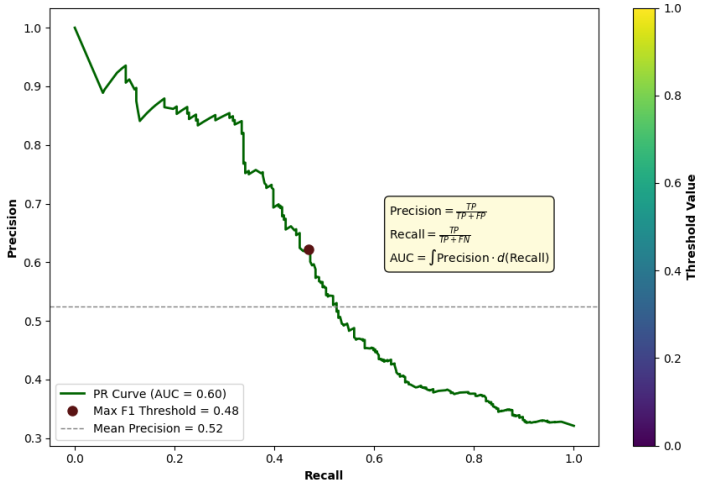


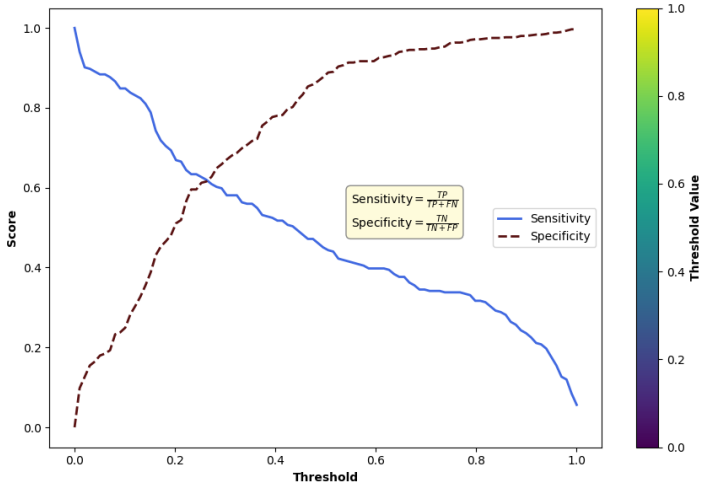
Figure 3. Decision tree representation



(a)



(b)



(c)

Figure 4. Classification model performance evaluation: (a) ROC curve with the AUC, (b) precision–recall curve with AUC and optimal threshold, and (c) sensitivity and specificity across thresholds

3.3.1. Mathematical foundation

The predictive model for student dropout is based on the RF algorithm, an ensemble learning method that combines multiple DTs to improve accuracy and robustness. The following mathematical formulations provide the theoretical foundation for this methodology [35].

– Data representation: we define the training dataset as:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \quad \mathbf{x}_i \in \mathbb{R}^d, \quad y_i \in \{0, 1\} \tag{1}$$

where $\mathbf{x}_i$ denotes the feature vector of student i, and y_i is the binary target variable: 1 if the student drops out, and 0 otherwise [36].

– Gini impurity: each DT splits the dataset using impurity functions. The Gini impurity is defined as [37]:

$$G(p) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2}$$

In binary classification ($K = 2$), it simplifies to:

$$G(p) = 2p(1 - p) \tag{3}$$

where p is the probability of belonging to one of the two classes (dropout or not).

– Shannon entropy (alternative): as an alternative to Gini, the Shannon entropy can be used:

$$H(p) = -\sum_{k=1}^{K} p_k \log_2(p_k) \tag{4}$$

– RF prediction: let $h_m(\mathbf{x})$ denote the prediction of tree m. The final prediction is based on majority voting:

$$\hat{y} = \text{mode}\left(h_1(\mathbf{x}),\ h_2(\mathbf{x}),\ \ldots,\ h_M(\mathbf{x})\right) \tag{5}$$

The estimated probability that a student drops out is:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(h_m(\mathbf{x}) = 1\right) \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise.

– Feature importance: the importance of each feature x_j is evaluated as:

$$\text{Imp}(x_j) = \sum_{t \in T_j} \frac{n_t}{n} \cdot \Delta\phi_t \tag{7}$$

where T_j is the set of nodes where feature x_j is used, n_t is the number of samples at node t, and $\Delta\phi_t$ is the impurity reduction at that node.

– Evaluation metrics: the model is evaluated using the following standard classification metrics.

Precision:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

F1-score:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

These metrics help determine how well the model identifies students at risk of dropping out.

4. RESULTS

In the results stage, Figure 5 illustrates the performance of the classification models—RF, XGBoost, and KNN over ten epochs, revealing distinct performance patterns. In Figure 5(a), RF consistently maintains the highest accuracy across training epochs, while KNN exhibits the lowest and most unstable accuracy, showing the stability and trend of each model during training. Figure 5(b), the precision metric follows a similar pattern, with RF and XGBoost achieving high values and KNN remaining low, highlighting how each algorithm's precision improves or fluctuates during training. Figure 5(c) shows that XGBoost attains the best recall over epochs, indicating strong performance in correctly identifying positive cases, whereas KNN performs poorly. Finally, Figure 5(d) confirms through the F1-score that XGBoost achieves the best balance between precision and recall throughout the epochs, followed by RF, while KNN continues to show the weakest performance across all metrics, reflecting the overall trade-off between precision and recall for each algorithm.

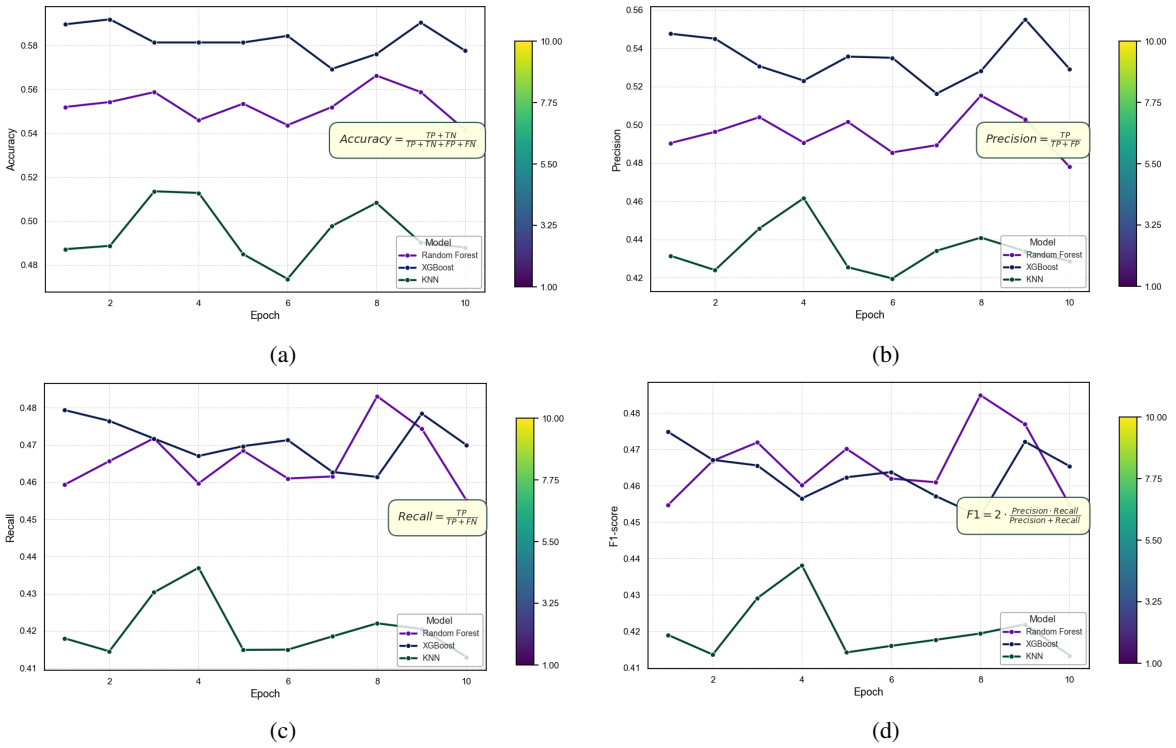


Figure 5. Algorithm comparison across epochs: (a) accuracy, (b) precision, (c) recall, and (d) F1-score

Table 7 shows that, in the classification problem addressed, the ensemble models RF and XGBoost consistently outperform KNN, with RF leading across all key performance metrics (accuracy, precision, recall, F1-score, and AUC), indicating its superior predictive reliability. Additionally, Table 8 highlights the feature importance analysis, emphasizing the critical role of GDP, unemployment rate, and mother's occupation as the most influential factors in the model's predictions—underscoring the significance of socioeconomic and macroeconomic variables. Finally, Table 9 presents the specific hyperparameter configurations for each model, which are essential for reproducibility and for understanding tuning process that optimized their performance.

Table 7. Performance comparison between classification algorithms

Modelo	Accuracy	Precision	Recall	F1-score	AUC
RF	0.87	0.86	0.85	0.85	0.91
XGBoost	0.85	0.84	0.83	0.83	0.89
KNN	0.76	0.73	0.71	0.72	0.76

Table 8. Feature importance for RF

Feature	Importance
GDP	0.25
Unemployment rate	0.18
Mother's occupation	0.12
Inflation rate	0.09
Tuition fees up to date	0.08

Table 9. Hyperparameters used by each model

Model	n_estimators	max_depth	learning_rate	n_neighbors
RF	100	10	0.0	0.0
XGBoost	150	5	0.1	0.0
KNN	0.0	0.0	0.0	5

5. DISCUSSION

Our results, as presented in Figures 4 and 5, align with and complement the existing literature on the application of machine learning in education. This includes both evaluating teaching performance and predicting student dropout. Our findings demonstrate that the proposed model—identified in Figure 5 —consistently outperforms algorithms such as RF, XGBoost, and KNN across metrics including accuracy, precision, recall, and F1-score, validating the effectiveness of robust classification approaches in this domain.

Specifically, the strength of our model in the ROC and precision-recall curves (Figure 5), with a competitive AUC (e.g., 0.909), directly contributes to the field of dropout prediction. It is comparable to the promising results reported by [13] and complements the work of [11], [12] by providing a solid foundation for predictive systems. While Vives *et al.* [14] achieved higher accuracy rates using LSTM networks, our comparison with commonly used machine learning models, including RF and XGBoost (Figure 5) is crucial, as our model surpasses these algorithms—aligning with findings from [15], [16], [21], [22], who advocate for the optimization and use of boosting algorithms to enhance predictive performance. Although our model does not employ hybrid or complex neural network architectures such as those explored by [17], [18]—who achieved extremely high accuracies (e.g., 98.1% with perceptron)—its consistently strong performance suggests a significant advancement among conventional classifiers for our specific task.

It is important to note that other studies, such as those by [19], [20], have reported even higher metrics (e.g., an AUC of 96.10% or accuracy of 0.963 using techniques like PCA and GBM), indicating the potential to enhance our model through advanced preprocessing methods or more complex architectures in future work. The importance of threshold adjustment, evident in our precision-recall curve, also resonates with the contributions of [23], [24]. The effectiveness of RF one of the base algorithms in our study—is further supported by research from [25], [26], who successfully applied it to dropout prediction in MOOCs and university programs, achieving strong metrics that validate its predictive capability in educational contexts. In this regard, the demonstrated capacity of our model, along with the other algorithms explored, highlights its high potential for integration into decision support systems and the formulation of preventive educational policies, as suggested by [27]. Overall, the results confirm the usefulness of machine learning in education and introduce an approach that improves existing models, offering practical support for decision-making in academic settings.

6. CONCLUSION

Student dropout in higher education is a multifaceted problem influenced by personal, academic, and socioeconomic factors, with significant consequences for institutions and society. To address this issue, a predictive model based on machine learning was developed using the RF algorithm. The model was trained on a CSV dataset obtained from Kaggle, containing variables related to university student dropout. The study followed the KDD methodology, encompassing data selection, preprocessing, transformation, data mining, and evaluation. During preprocessing, data inconsistencies and outliers were removed, and the most relevant variables were selected. Exploratory statistical analyses were conducted to assess data quality and structure. In the data mining phase, predictive models were built and evaluated using performance metrics such as the ROC curve, precision–recall curve, AUC, and maximum F1-score. A comparative analysis was performed among

RF, XGBoost, and KNN models using accuracy, precision, recall, and F1-score across training iterations. The results indicate that the RF-based model achieved superior performance, demonstrating high accuracy and reliability in predicting student dropout. Despite these promising findings, the study is limited by the use of a single static dataset without external validation. Future research should incorporate real-world and diverse datasets to improve generalizability and explore system-level implementations using REST APIs and data storage architectures such as data warehouses.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Laberiano Andrade-Arenas	✓	✓				✓			✓			✓	✓	
Inoc Rubio Paucar	✓	✓	✓	✓	✓	✓			✓	✓	✓			
Margarita Giraldo Retuerto	✓	✓	✓		✓	✓			✓	✓				
Cesar Yactayo-Arias	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.

REFERENCES

[1] G. B. N. Silva, "Student dropout in higher technological institutes of Ecuador: a review of the literature (in Spanish: Deserción estudiantil en Institutos Superiores Tecnológicos de Ecuador: Una revisión de la literatura)," *Revista Latinoamericana Ogmios*, vol. 3, no. 8, pp. 25–32, Jun. 2023, doi: 10.53595/rlo.v3.i8.074.

[2] M. E. L. Rivera and J. M. R. Vásquez, "The economic factor as the main cause of university student dropout in Central America," *Entorno*, vol. 1, no. 74, pp. 60–70, Sep. 2023, doi: 10.5377/entorno.v1i74.15668.

[3] A. A. Andrade and L. C. K. Salinas, "Causes of university students desertion in pandemic times at the Bolivian Catholic University," *Revista Educación*, Jan. 2024, doi: 10.15517/revedu.v48i1.56040.

[4] A. Kuz and R. Morales, "Educational data science and machine learning: a case study on university student dropout in Mexico (in Spanish: Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México)," *Education in the Knowledge Society*, vol. 24, Jun. 2023, doi: 10.14201/eks.30080.

[5] C. R. Parra, N. C. Manjarrez, E. J. O. González, and Y. V. Pérez, "Factors influencing student attrition intertiary education in Colombia," *Revista Interdisciplinaria de Humanidades, Educación, Ciencia y Tecnología*, vol. 9, no. 17, pp. 45–56, Jul. 2023, doi: 10.35381/cm.v9i17.1122.

[6] J. M. C. Matute, A. V. Franco, and J. I. T. Segarra, "Factors that influence student desertion in the academic unit of social sciences of the Catholic University of Cuenca," *Conciencia Digital*, vol. 6, no. 3, pp. 30–48, Jul. 2023, doi: 10.33262/concienciadigital.v6i3.2621.

[7] S. M. D. L. F. Valdez and Y. L. Lara, "Desertion and student reprobation, campus mederos, UANL (in Spanish: Deserción y reprobación estudiantil, campus mederos, UANL)," *Multidisciplinas de la Ingeniería*, vol. 7, no. 10, pp. 1–11, Dec. 2023, doi: 10.29105/mdi.v7i10.212.

[8] L. F. C. Rojas, E. E. Peña, and E. R. Cuero, "Analysis of characteristics influencing student dropout in the context of a Latin American university (in Spanish: Análisis de características que influyen en la deserción estudiantil en el contexto de una universidad latinoamericana)," *Revista EIA*, vol. 20, no. 40, Dec. 2023, doi: 10.24050/reia.v20i40.1628.

[9] H. V. Torres, J. Á. Morales, W. M. Rodriguez, and J. C. Mejía, "Classification model for student dropout in public universities in Peru," *Revista de Ciencias Sociales*, vol. 30, no. 1, pp. 452–469, Feb. 2024, doi: 10.31876/rcs.v30i1.41667.
[10] F. O. C. Guashpa, L. J. G. Ruiz, W. V. P. Quiroz, and R. M. B. Chila, "Student desertion from the automotive mechanics career at the instituto superior tecnológico Luis Tello (in Spanish: Deserción Estudiantil de la Carrera de Mecánica Automotriz del Instituto Superior Tecnológico Luis Tello)," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 7, no. 5, pp. 4502–4516, 2023, doi: 10.37811/cl_rcm.v7i5.8054.
[11] R. Ahuja and S. C. Sharma, "Stacking and voting ensemble methods fusion to evaluate instructor performance in higher education," *International Journal of Information Technology*, vol. 13, no. 5, pp. 1721–1731, Oct. 2021, doi: 10.1007/s41870-021-00729-4.
[12] D. A. G. Pachas, G. G. Zanabria, E. C. Vargas, G. C. Chavez, and E. G. Nieto, "Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through machine learning and survival analysis methods in the Latin American context," *Education Sciences*, vol. 13, no. 2, Feb. 2023, doi: 10.3390/educsci13020154.
[13] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, 2022, doi: 10.1016/j.caeai.2022.100066.
[14] L. Vives *et al.*, "Prediction of students' academic performance in the programming fundamentals course using long short-term memory neural networks," *IEEE Access*, vol. 12, pp. 5882–5898, 2024, doi: 10.1109/ACCESS.2024.3350169.
[15] C. L. R. Velasco, E. G. Villena, J. B. Ballester, F. Á. D. Prados, E. S. Alvarado, and J. C. Álvarez, "Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 04, pp. 120–155, Feb. 2023, doi: 10.3991/ijet.v18i04.34825.
[16] C. H. Cho, Y. W. Yu, and H. G. Kim, "A study on dropout prediction for university students using machine learning," *Applied Sciences*, vol. 13, no. 21, Nov. 2023, doi: 10.3390/app132112004.
[17] A. C. T. Mary and A. L. P. J. Rose "Ensemble machine learning model for university students' risk prediction and assessment of cognitive learning outcomes," *International Journal of Information and Education Technology*, vol. 13, no. 6, pp. 948–958, 2023, doi: 10.18178/ijiet.2023.13.6.1891.
[18] N. Mouchantaf and M. Chamoun, "Predicting student dropout with minimal information," *Iraqi Journal of Science*, pp. 5265–5279, Oct. 2023, doi: 10.24996/ijs.2023.64.10.33.
[19] S. Kim, E. Choi, Y.-K. Jun, and S. Lee, "Student dropout prediction for university with high precision and recall," *Applied Sciences*, vol. 13, no. 10, May 2023, doi: 10.3390/app13106275.
[20] H. V. Torres, W. M. Rodriguez, J. Á. Morales, J. C. Mejía, and C. M. Murillo, "Classification model for student attrition in a Peru public university," *Salud, Ciencia y Tecnología - Serie de Conferencias*, vol. 2, May 2023, doi: 10.56294/sctconf2023175.
[21] D. E. M. D. Silva, E. J. S. Pires, A. Reis, P. B. D. M. Oliveira, and J. Barroso, "Forecasting students dropout: a UTAD university study," *Future Internet*, vol. 14, no. 3, Feb. 2022, doi: 10.3390/fi14030076.
[22] A. Villar and C. R. V. D. Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, vol. 4, no. 1, Jan. 2024, doi: 10.1007/s44163-023-00079-z.
[23] A. G. Nucamendi, J. Noguez, L. Neri, V. R. Rella, and R. M. G. G. Castelán, "Predictive analytics study to determine undergraduate students at risk of dropout," *Frontiers in Education*, vol. 8, Oct. 2023, doi: 10.3389/feduc.2023.1244686.
[24] J. O. Q. Quispe, O. C. Toledo, M. C. Toledo, E. E. C. Llatasi, and E. Saira, "Early prediction of university student dropout using machine learning models," *Nanotechnology Perceptions*, pp. 659–669, 2024.
[25] S. Dass, K. Gary, and J. Cunningham, "Predicting student dropout in self-paced MOOC course using random forest model," *Information*, vol. 12, no. 11, Nov. 2021, doi: 10.3390/info12110476.
[26] M. V. Martins, L. Baptista, J. Machado, and V. Realinho, "Multi-class phased prediction of academic performance and dropout in higher education," *Applied Sciences*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084702.
[27] A. J. F. Garcia, J. C. Preciado, F. Melchor, R. R. Echeverria, J. M. Conejero, and F. S. 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.
[28] M. Hinojosa, I. Derpich, M. Alfaro, D. Ruete, A. Caroca, and G. Gatica, "Student clustering procedure according to dropout risk to improve student management in higher education," *Texto Livre*, vol. 15, Mar. 2022, doi: 10.35699/1983-3652.2022.37275.
[29] A. K. H. Robles and E. D. V. Valle, "Labor conditions of urban students and school dropout at the upper secondary level in Mexico," *Estudios Demográficos y Urbanos*, vol. 31, no. 3, pp. 663–696, Sep. 2016, doi: 10.24201/edu.v31i3.1653.
[30] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
[31] S. Georganos *et al.*, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling," *Geocarto International*, vol. 36, no. 2, pp. 121–136, Jan. 2021, doi: 10.1080/10106049.2019.1595177.
[32] G. A. Romero, C. A. G. Prieto, M. A. D. Barriosnuevos, and N. A. R. Menjura, "Review and perspectives for the construction of robust databases with missing data: case applied to financial information," *Tecnura*, vol. 27, no. 75, pp. 12–37, Jan. 2023, doi: 10.14483/22487638.18268.
[33] C. V. López, E. T. Guerrero, C. S. Noroña, and C. F. Cadena, "Identification of significant variables in student dropout, using a KDD linear regression mathematical model," *Revista Científica Ciencia y Tecnología*, vol. 22, no. 36, Oct. 2022, doi: 10.47189/rcct.v22i36.493.
[34] A. Shamim, "Student dropout & success prediction dataset," *Kaggle*. Accessed: May 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/adilshamim8/predict-students-dropout-and-academic-success
[35] A. Shahin, A. S. Aboumasoudi, B. Khamoushpour, and S. Khademolqorani, "Measuring and predicting the service quality of information systems and technology: an integrated approach of decision tree and random forest," *International Journal of Business Performance Management*, vol. 1, no. 1, 2025, doi: 10.1504/IJBPM.2025.10058530.
[36] H.-H. Zou *et al.*, "Rapid detection of colored and colorless macro- and micro-plastics in complex environment via near-infrared spectroscopy and machine learning," *Journal of Environmental Sciences*, vol. 147, pp. 512–522, Jan. 2025, doi: 10.1016/j.jes.2023.12.004.

[37] J.-M. Cao *et al*., "Predicting the efficiency of arsenic immobilization in soils by biochar using machine learning," *Journal of Environmental Sciences*, vol. 147, pp. 259–267, Jan. 2025, doi: 10.1016/j.jes.2023.11.016.

BIOGRAPHIES OF AUTHORS

Laberiano Andrade-Arenas is a doctor in systems and computer engineering. Master in Systems Engineering. Graduated with a master's degree in University Teaching. Graduated with a master's degree in accreditation and evaluation of educational quality. Systems Engineer, scrum fundamentals certified, a research professor with publications in Scopus-indexed journals. He can be contacted at email: landrade@uch.edu.pe.

Inoc Rubio Paucar bachelor in Systems and Computer Engineering. He has a background in database management and computer system design, with a focus on artificial intelligence applications, machine learning, and data science. His research interests are in the area of computer science. He can be contacted at email: Enoc.Rubio06@hotmail.com.

Margarita Giraldo Retuerto is a Systems Engineer with a solid background in software engineering, specialized in the analysis, design, and development of technological solutions. She has experience applying software development methodologies and quality best practices. She shows a strong interest in scientific research, particularly in areas related to emerging technologies and process improvement. She has contributed to academic and technical projects with an analytical and methodical approach. Her professional profile combines technical expertise, critical thinking, and a strong research orientation. She can be contacted at email: mgiraldo@uch.edu.pe.

Cesar Yactayo-Arias is bachelor's degree in administration from Universidad Inca Garcilazo de la Vega and a master's degree in education from Universidad Nacional de Educación Enrique Guzmán y Valle, he is a doctoral candidate in administration at Universidad Nacional Federico Villarreal. Since 2016 he has been teaching administration and mathematics subjects at the Universidad de Ciencias y Humanidades and since 2021 at the Universidad Continental. Currently, he also works as an administrator of educational services at the higher level, he is the author and co-author of several refereed articles in journals, and his research focuses on TIC applications to education, as well as management using computer science and the internet. He can be contacted at email: yactayocesar@gmail.com.