ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.571

Predictive Model to Reduce Undergraduate Student Dropout at the Army Scientific and Technological Institute of Peru

Víctor Mariscal Carhuamaca, Msc[1], Carlos Quinto Huamán, PhD[2], Gladys Madeleine Rojas Cangahuala, PhD[1], Patricia Fernández Muriel, Msc[1], and Juan Godoy Caso, PhD[1]
[1]Grupo de Investigación en Ciberseguridad, IoT e Inteligencia Artificial (GriCIA), Instituto Científico y Tecnológico del Ejército Lima, Perú, vmariscalc@icte.edu.pe, grojasc@icte.edu.pe, pfernandezm@icte.edu.pe, jgodoyc@icte.edu.pe
[2]Universidad Privada del Norte, Lima, Perú, carlos.quinto@upn.pe

Abstract*– Student dropout represents a problem of great complexity and repercussion in the university educational environment, affecting students, academic institutions and society. The Army Scientific and Technological Institute of Peru (ICTE) is not exempt from this problem, in addition to the small amount of historical data and limited access to information due to personal data protection issues. To address this problem, this paper presents a predictive model to reduce dropout in undergraduate students at the ICTE. It uses Machine Learning techniques and compares its prediction levels through reliable performance metrics most used in the literature. The dataset initially concentrated information from 144 students, classified into personal data, socioeconomic factors and academic performance. To overcome this challenge, we chose to generate synthetic data using the SMOTE technique based on the original dataset, thus facilitating the training of the classification algorithms, balancing minority classes and reducing biases in the prediction. The results obtained highlight the exceptional performance of the LightGBM model, which achieved 95% accuracy for training and testing. This model provides ICTE a strategic tool for implementing preventive measures to mitigate factors that negatively affect or hinder student retention in academic training. This approach promises to be beneficial for students and the institution, contributing significantly to educational quality.

Keywords*-- Prediction model, predictive analytics, dropout factors, lightGBM model, student dropout, academic performance.

I. Introduction

Student dropout in education represents a highly complex issue that adversely impacts the learning process, with negative repercussions that affect students, academic institutions, economic resources, and society [1], [2], [3]. It is a significant problem for industrialized and developing countries, but its impact is even more accentuated in less developed economies, with a significant impact on the training of future professionals and on the prestige of educational institutions [4], [5]. In Peru, the Army Scientific and Technological Institute, in its commitment to provide quality education, faces the need for early identification of students at risk of academic dropout. Early identification of students at risk is essential to provide adequate support and take preventive measures to enable students to overcome academic difficulties [6].

The literature describes several variables that may influence the academic performance of college students, including academic background, socioeconomic factors, participation in extracurricular activities, and study skills. While these factors have been studied in depth, accurate prediction of dropout risk at an early stage remains a challenge [7]. It also highlights the importance of clearly defining the space and time of the students' study context, as this delineation is essential to identify effective and demonstrable solutions. Given the risk that student dropout represents for the university sector, the relevance of undertaking efforts to anticipate and mitigate this phenomenon is emphasized, bringing benefits to both educational institutions and the sector as a whole [8], [9]. In this sense, the main objective of this paper is to implement a predictive model to identify the factors that influence the dropout of university students of the Army Scientific and Technological Institute.

The paper is structured as follows: Section 2 presents a comprehensive review of the literature related to the research topic. Section 3 describes the data set used. Section 4 presents the proposed predictive model. The results obtained and discussion are presented in Section 5. Finally, Section 6 describes the main conclusions of the research.

II. Overview of Machine Learning Strategies to Reduce Student Dropouts

After reviewing the diversity of information on the dropout of university students, a wide range of solutions have been identified that make it possible to anticipate the probability of student dropout. In this context, theoretical models, study variables, parameters and algorithms adapted to various situations and contexts have been proposed. The scopes and approaches of these works vary significantly according to the period or demographic space and according to the specific solutions for each institution. In [10], the authors propose the construction of a model for early prediction of students at risk of dropping out of college. They use the Logistic Regression algorithm adding a regularization term and the Input-Output Hidden Markov Model (IOHMM). The results showed that the proposed models achieved 84 % accuracy compared to the baseline machine learning models for predicting students at risk of dropping out. In [1], they considered aspects to study modality, training, testing strategy,

cross-validation and confusion matrix. The results of the review revealed that the most used algorithm was Random Forest, present in 21.73% of the studies; this algorithm obtained an accuracy of 99% in predicting student dropout, higher than all the algorithms used in the total number of studies reviewed. In [3], they present an early warning system to automatically identify first semester students at high risk of dropping out. The system is based on a machine learning model trained from historical data of first semester students. The results show that the system can predict students at risk, with a sensitivity of 61.97 %, allowing early intervention for those students and reducing the dropout rate. In [11], the J48 WEKA supervised machine learning algorithm was developed using decision trees, thus identifying the main factors that influenced dropout, with the J48 algorithm an accuracy of 87.76% and 66.63% of moderate agreement was obtained through Cohen's Kappa index, concluding that based on the results obtained, managers can identify students with possible risk of dropping out and thus take corrective measures and implement strategies to help reduce the student dropout rate. In [12], a Naïve Bayes classification and a Support Vector Machine (SVM) algorithm were presented to predict academic performance of graduate students using demographic factors and first semester exam scores, achieving outstanding results. In [13], the FWTS-CNN dropout prediction model is proposed, which achieved more than 87% accuracy, an improvement of about 2% overusing the CNN algorithm. The FWTS-CNN model integrates the effects of behavioral characteristics and behavioral time on dropout, which effectively improves the dropout prediction accuracy. In [14], a predictive model based solely on academic data was implemented. The models known as Gradient Boosting, Random Forest and Support Vector Machine were used, achieving satisfactory results. At the end of the research, the authors conclude that predictive models serve as inputs for a decision support system and can be used to promote effective dropout prevention policies. In [15], the authors ensure that predictive capacity is defined as the ability to develop and evaluate models aimed at generating predictions about new observations, making use of advanced statistical tools for the purpose of anticipating future events. To establish this predictive capability, organizations must rely on a predictive analytics platform that integrates data warehouses, predictive analytics algorithms, dashboards, and reporting systems that facilitate optimal decision making for users.

A. *Variables for predicting student* dropout

Several factors have been identified to predict university student dropout [16]. Table 1 shows some relevant variables obtained from an exhaustive literature review. These elements allow the underlying causes of student dropout to be analyzed from sociological, economic, and organizational perspectives [17]. It is essential to address these dimensions to obtain a holistic understanding of the reasons that lead to dropout.

TABLE I
MAIN DROPOUT VARIABLES USED IN THE LITERATURE

VARIABLE	DESCRIPTION	REFERENCE
Age	Age of student	[18], [19], [20]
Gender	Gender of Student	[18], [19], [20]
Marital Status	Marital Status of Student	[19], [21]
Employment	If the student is employed	[19], [22], [21]
Athlete	If the student practices any sport	[22]
Place of Residence	City or district where the student lives	[19]
Own home	If the home is owned by the family	[18]
Social Status	Related to socioeconomic level	[18], [19]
Transportation	Primary means of transportation used by the student	[22]
Number of siblings	Number of siblings of the student	[19], [22]
Academic level of father	Academic level of the student's father	[19], [22]
Academic level of mother	Academic level of the student's mother	[19], [22]
Scholarship	If the student is in receipt of a scholarship	[22], [23]
Entrance qualification	Student's entrance or admission qualification	[23]
I semester qualification	Average qualification of the first semester	[2], [22]
II semester qualification	Average qualification of the second semester	[2], [22]
University subjects failed in the 1st semester	Number of subjects failed by the student in the 1st Semester	[2], [22]
University subjects failed in the 2nd semester	Number of subjects failed by the student in the 2nd Semester	[2], [19]
Dropout	Attribute or target class, information about the student's current situation.	[20]

B. *Factors influencing dropout prediction*

The academic performance of university students is a crucial topic in educational research. The literature highlights several determinants for predicting student dropout, including prior academic preparation, study skills, and the learning environment. The complex interaction of these elements highlights the need for a comprehensive analysis to ensure dropout prevention. This in-depth knowledge will enable the effective design of retention strategies, promoting student success and retention in higher education [21].

C. *Prediction models*

In [24], a prediction model capable of anticipating a student's risk of dropping out is proposed, using information from the first three semesters taken by students of the bachelor's degree in computer science. Currently, Educational Management Systems stores a vast amount of data from

interactions not only between students and teachers, but also between students and their educational environment. Analyzing and discovering patterns manually from such a large volume of data is a complex task, which has led to the widespread use of educational data mining. In this study, the CRISP-DM methodology is applied and data from undergraduate students in Computer Science at the Federal University of Pelotas, Brazil are used. The results of three algorithms are presented: the Decision Tree algorithm yields an accuracy of 84.80%, a Recall of 85.80% and an area under the curve (AUC) of 77.24%; the Random Forest algorithm achieves an accuracy of 88.57%, a Recall of 90.14% and an AUC of 83.22%; while the Logistic Regression algorithm obtains an accuracy of 71.24%, a Recall of 94.28% and an AUC of 58.39%. These results suggest that it is feasible to use a predictive model based exclusively on data from the first three semesters of the academic program. In [25], an approach to predict the likelihood of student dropout is proposed through a hybrid model. This hybrid model incorporates a dual-channel convolutional neural network (CNN) to automatically identify relevant features from student progress records. Then, it employs an attention mechanism to highlight crucial information. Finally, a Temporal Convolutional Network (TCN) is applied to capture the relationships between these hidden features at different time scales. The results of comprehensive experiments on the KDD CUP 2015 dataset demonstrate that the proposed model outperforms other dropout prediction methods in terms of performance. In [26], a predictive model is presented with the objective of reducing the dropout rate of university students in Peru. This model is composed of three phases of predictive analysis that are integrated with the stages defined by the IBM SPSS Modeler methodology. Bayesian network techniques were evaluated in comparison with decision trees due to their superior level of accuracy relative to other algorithms in the context of educational data mining. Data were collected from 500 undergraduate students belonging to a private university in Lima. The results indicate that Bayesian networks outperform decision trees in terms of metrics such as precision, accuracy, specificity, and error rate. Specifically, in a training sample with a ratio of 8:2, the accuracy of Bayesian networks reaches 67.10%, while the accuracy of decision trees is 61.92%. Furthermore, it is observed that the variables "sports person" (0.30%), "own house" (0.21%) and "high school grades" (0.13%) have the greatest influence on the prediction model, both for Bayesian networks and decision trees. In [27], a predictive analysis model was implemented to identify students at risk of dropping out of Peruvian universities and the variables that influence this. For this purpose, the Cross Industry Standard Process for Data Mining (CRISP - DM) methodology is used to develop the model and four Machine Learning algorithms. The methodology consists of five phases: business understanding, data understanding, data preparation, modeling, and evaluation. The experiment was carried out by conducting a survey to 385 students from different public and private universities in Peru, where cognitive, affective, family environment, pre-university, career, and university variables were considered. The results showed that the most influential variables in the prediction of university dropout were the student's "age", "term" and "method of financing". We also found that the Random Forest algorithm performed the best, with an AUC of 0.9623 in predicting college dropout.

III. DESCRIPTION OF DATASET

For this work, data were collected from 144 undergraduate students of the Army Scientific and Technological Institute during two consecutive academic periods (semesters). To preserve privacy and simplify the collection of anonymized data, we chose to work with synthetic data generated from this initial sample. The main objective was to expand this original sample to a data set of 1440 students, thus speeding up the data collection process and guaranteeing the integrity and privacy of the information.

A. Personal and academic data characterization

Socioeconomic and academic data were obtained from undergraduate students during the academic periods between 2022 and 2023. At an initial stage of the research, data was collected from one hundred forty-four (144) students, including nineteen (19) relevant attributes. It is crucial to highlight that this initial sample contains confidential information from student enrollment forms, as well as data generated during their academic trajectory. The confidentiality of the students was maintained by identifying them using whole numbers, excluding any identifying information such as names, codes or identification numbers.

The attribute 'Dropout' acts as the class field in our study, providing crucial information about the academic status of students. According to the definition of the Peruvian Ministry of Education, a student is considered a dropout if he/she abandons his/her studies for two consecutive academic periods without making an enrollment. This attribute is fundamental for understanding and predicting student dropout in our research. Table 2 presents the coded variables, considering their suitability for the context of the model. The choice of this coding was based on the necessity to adapt the model to the limitations of access to certain confidential data, thus ensuring integrity and compliance with privacy regulations.

TABLE II
VARIABLES CODED FOR THE PROPOSED MODEL

VARIABLE	CATEGORY	DOMAIN
Age	Numerical Range	[0, N]
Gender	[M, F]	[0, 1]
Marital Status	[Single, Married, Widowed, Divorced]	[0, 1, 2, 3, 4]
Employment	[Yes, No]	[0, 1]
Athlete	[Yes, No]	[0, 1]
Place of Residence	[Alphabetically the 43 Districts of Lima]	[0…43]
Own home	[Yes, No]	[0, 1]

Social Status	[High, Medium High, Medium, Medium Low, Low]	[1, 2, 3, 4, 5]
Transportation	[Yes, No]	[0, 1]
Number of siblings	Numerical Range	[0, N]
Academic level of father	[Primary, Secondary, High School, University, None]	[1, 2, 3, 4]
Academic level of mother	[Primary, Secondary, High School, University, None]	[1, 2, 3, 4]
Scholarship	[Yes, No]	[0, 1]
Entrance qualification	Numerical Range	[0, 20]
I semester qualification	Numerical Range	[0, 20]
II semester qualification	Numerical Range	[0, 20]
University subjects failed in the 1st semester	Numerical Range	[0, N]
University subjects failed in the 2nd semester	Numerical Range	[0, N]
Dropout	[Yes, No]	[0, 1]

B. Synthetic data

Synthetic data, created using machine learning techniques, are essential for preserving privacy, testing systems and training algorithms. Their generation responds to the growing need for specific information. These data allow exploring different tasks in data science and guarantee the anonymity of the generated samples [28], thus addressing privacy challenges in sensitive information. Synthetic data represents an effective strategy to generate datasets that simulate the features and structures of the original data. Table 3 describes the most used tools for generating synthetic data.

TABLE III
TOOLS TO GENERATE SYNTHETIC DATA

TOOL	DESCRIPTION
GPT-J	An open-source option that competes with OpenAI's GPT-3 text generation tool.
Synthea	A widely recognized open-source tool in the medical field.
Scikit-learn	Used to create synthetic data sets to support regression, clustering, and classification tasks to enable forecasting.
SymPy	Used by data science professionals who require highly customized synthetic datasets for specific needs, as it allows the creation and development of tailor-made symbolic expressions.
Pydbgen	Used to generate common data such as telephone numbers or e-mail addresses.
Synthpop	An R package used to produce synthetic demographic data.
Faker	A Python package that generates synthetic data, such as names, addresses, e-mail addresses, Social Security numbers and other types of information.
Synthetic Data Vault (SDV)	A Python library used to generate tables, relational databases, and time series models.

C. Synthetic data for the proposal

Synthetic data was generated from the initial dataset of 144 records to 1440 instances. The Google Colab platform and the HMASynthesizer method of Synthetic Data Vault (SDV) were used. Figure 1 shows the correlation of the synthetic data, describing how the procedures were carried out and how the generated data are related.

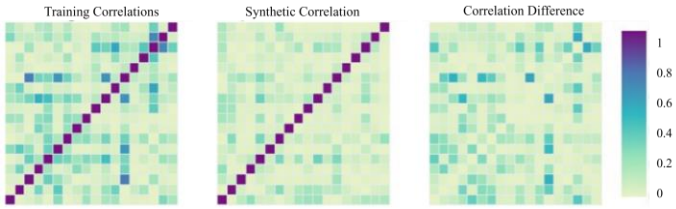

Fig. 1 Correlation of synthetic data

After creating the synthetic data, a significant imbalance between classes within the synthetic dataset became apparent. This dataset can generate biases and affect the effectiveness of machine learning models by favouring the majority class and under-representing the minority class. To address this issue, we implemented the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic samples of the minority class by interpolating existing instances. This strategy allows increasing the representation of the least frequent class, thus balancing the distribution of classes in the SMOTE dataset [29].

IV. PROPOSED PREDICTIVE MODEL

The proposed predictive model aims to detect and reduce the risk of undergraduate student dropout. The process was divided into two stages: (i) training, to prepare and process the data; and (ii) testing, to perform the predictive analysis using classification algorithms such as Voting Classifier, Gradient Boosting, eXtreme Gradient Boosting (XGBoost) and LightGBM (See Figure 2).

A. Training Stage

A.1. Data preparation

Academic performance, socio-economic and personal data of 144 students have been collected and extracted from various data sources of the Academic Department of the Army Science and Technology Institute.

A.2. Data processing

The dataset of 144 students was coded, and data cleaning, outlier removal and enrichment through synthetic data generation was performed. This process expanded the dataset to a total of 1440 rows. The main purpose of this enrichment was to prepare a dataset sufficiently robust for implementation in classification algorithms. The dataset was then segmented into training data (80%) and test data (20%).

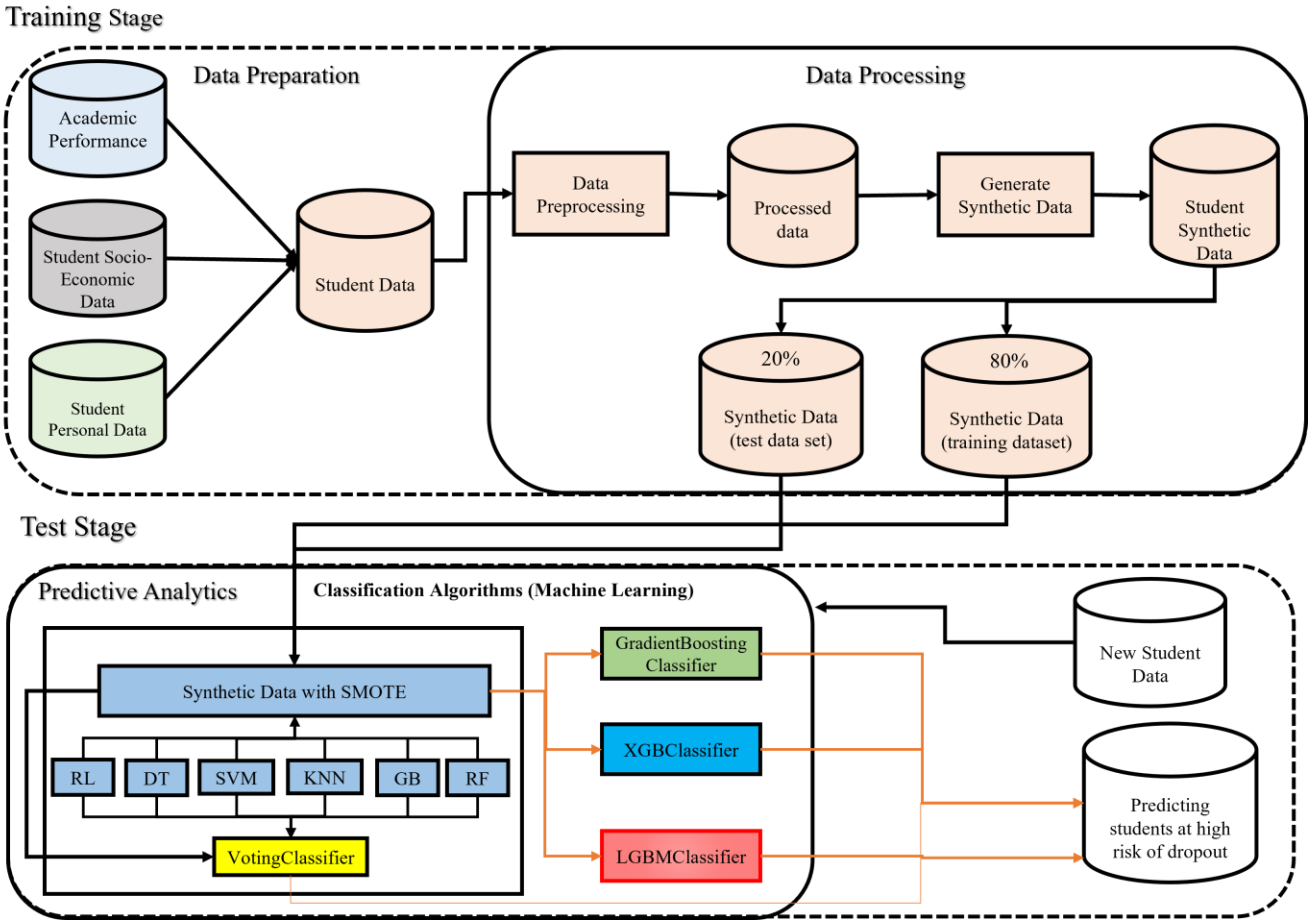

Fig. 2 Proposed Model Architecture

B. *Test Stage*

At this stage we identified that the synthetic dataset exhibited a noticeable imbalance in the training and test set. In response to this challenge, the SMOTE class balancing technique was employed.

The Voting Classifier prediction model was used to construct a classifier. In addition, a comparative evaluation of Gradient Boosting, eXtreme Gradient Boosting (XGBoost) and LightGBM algorithms was performed to identify and choose the most suitable algorithm for the dataset. This comparison was carried out to determine the model that best suits our particularities, thus ensuring accurate and reliable predictions in the context of our study.

B.1. Voting Classifier

We use this algorithm to improve the precision of a prediction model by combining two or more algorithms. Initially, training sets are fed to different algorithms to build a classification model (See Figure 3). Finally, based on the majority vote, the classification model assigns a label to the test data sets. For this reason, the voting classifier is also called an ensemble classification model [30].

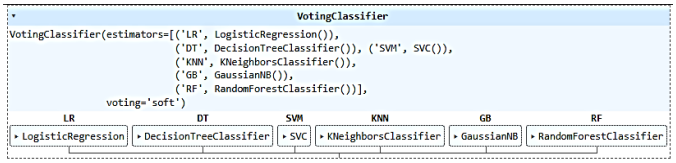

Fig. 3 Voting Classifier

The Voting Classifier prediction model starts when the synthetic training dataset is assigned to different classifiers, as shown in Figure 4, including Logistic Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, Naive Bayes, and Randon Forest. These algorithms generate class labels for unknown samples by splitting the data set into training and testing. The results are delivered to the voting classifier, which, based on the majority, assigns class labels to the known data.

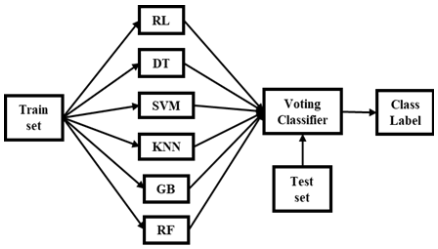

Fig. 4 Voiding Classifier Architecture

B.2 Gradient Boosting

This classification algorithm generates a model that is built additively in different stages, allowing the optimization of various loss functions that are differentiable. In each stage n_classes_ regression trees are fit on the negative gradient of the loss function, e.g. binary or multiclass log loss. Binary classification is a special case where only a single regression tree is induced.

B.3. eXtreme Gradient Boosting (XGBoost)

This algorithm stands out for its scalability and its ability to efficiently handle data sparsity. It considers compression, fragmentation and optimizes cache access [32]. In this approach, each tree is trained using the residual error of the previously constructed tree, leading to improvements in the overall model performance. The final prediction is obtained by summing the individual predictions of each tree built in the process [33].

B4. LightGBM (LGM)

It is a decision tree learning algorithm based on gradient boosting that has been widely used in feature selection, classification, and regression [34]. This method uses strategies such as Gradient-Based One-Side Sampling (GOSS) to discern between samples with different gradients, prioritizing those with larger gradients and randomly sampling samples with smaller gradients. This technique drastically reduces the computations required during model training, improving computational efficiency. In addition, LightGBM implements Exclusive Feature Bundling (EFB), which groups exclusive features into histograms to form more compact feature sets, reducing complexity and training time. In addition, it optimizes row and column subsampling to intelligently select relevant samples and features, which speeds up the training process and improves prediction accuracy [35].

V. RESULTS AND DISCUSSION

A. Results

This section presents the results obtained from the processing and predictive analysis of the data. Figure 5 shows the distribution of one thousand four hundred forty (1440) synthetic data generated from the initial set. Zero "0" is the non-dropout student and a "1" is the dropout student.

Two prediction scenarios were carried out: (i) without using SMOTE technique and (ii) using SMOTE technique. In both cases the Random Forest (RF) classification algorithm was applied to evaluate the prediction precision. Figure 6 shows the confusion matrix of the experiment without SMOTE. It is evident that the class imbalance in the dataset has led the model to classify a greater number of instances as non-dropout students. Table 4 shows the performance of the random forest algorithm without using SMOTE. The performance metrics precision, recall and F1_Score reach values close to 1 for the class "non-dropout student"; this result confirms that the classification model learns further about the majority class, demonstrating that it is necessary to balance the classes to achieve efficient results.

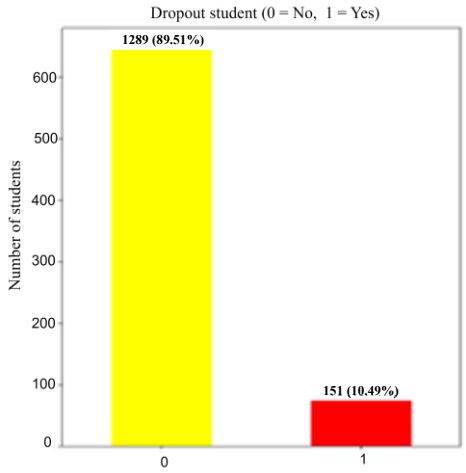

Fig. 5 Synthetic data distribution

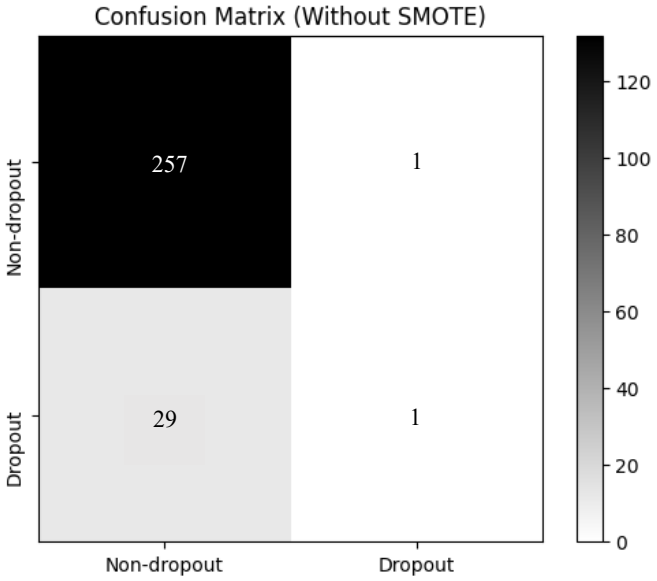

Fig. 6 Confusion Matrix (Without SMOTE)

TABLE IV
RF CLASSIFICATION METRICS (WITHOUT SMOTE)

	precision	recall	f1_score	support
Non-dropout	0.90	1.00	0.94	258
Dropout	0.50	0.03	0.06	30
accuracy			0.90	288
macro avg	0.70	0.51	0.50	288
weighted avg	0.86	0.90	0.85	288

Figure 7 shows the confusion matrix using SMOTE, while Table 5 presents the detailed report of the Classification

model with the inclusion of the oversampling technique. The results obtained indicate that the class balancing of the dataset improves the predictive ability of the algorithm. The three metrics used achieve more realistic values above 0.92, which ensures a classification with low probability of bias between classes.

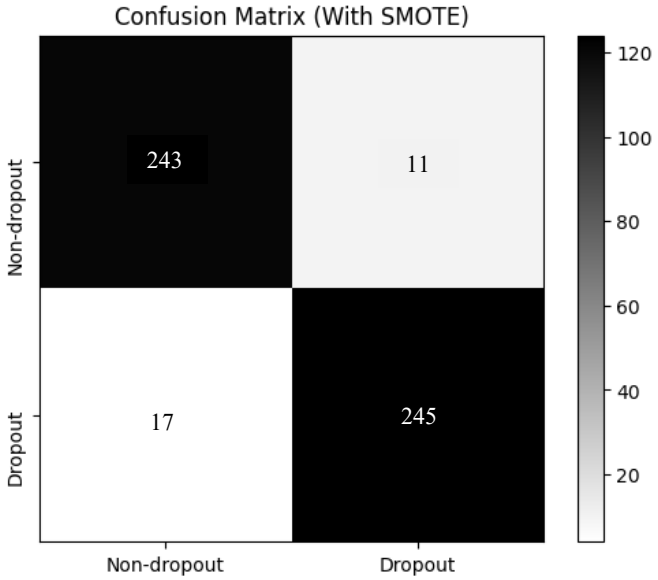

Fig. 7 Confusion Matrix (With SMOTE)

TABLE V
RF CLASSIFICATION METRICS (WITH SMOTE)

	precision	recall	f1_score	support
Non-dropout	0.93	0.96	0.95	254
Dropout	0.96	0.94	0.95	262
accuracy			0.95	516
macro avg	0.95	0.95	0.95	516
weighted avg	0.95	0.95	0.95	516

After applying the SMOTE technique, various machine learning algorithms were employed. Each classification algorithm underwent training and evaluation using performance metrics such as the confusion matrix, accuracy, recall, and F1-score. The confusion matrix describes the classification model's performance when assessing a set of test data with known true values. This matrix organizes the model's predictions into four categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Accuracy represents the proportion of correct predictions made by the classifier in relation to the total predictions. Simply put, this metric indicates how frequently the classifier makes accurate predictions. It is calculated by dividing the number of correct predictions by the total predictions made by the model, offering a measure of the classifier's accuracy in classifying the samples. The recall, indicating the rate of true positives, represents the ratio of correct positive predictions to the total number of true positive instances. This metric assesses the classifier's effectiveness in detecting all positive instances present in the dataset. F1-Score represents the harmonic measure between recall and accuracy, where a higher value indicates better model performance. This metric considers both the model's ability to correctly identify relevant samples and its accuracy in predicting the target class accurately. A higher F1 score reflects an optimal balance between the model's precision and recall, indicating a better overall classification capability.

Table 6 shows the performance of six classification algorithms within the Voting Classifier model. Notably, the Random Forests algorithm stands out with an impressive accuracy of 0.95, recall of 0.93, and an F1-Score of 0.95.

TABLE VI
PERFORMANCE OF VOTING CLASSIFIER ALGORITHMS

Algorithm Classifier	accuracy-score	recall-score	f1_Score
Logistic Regression (LR)	0.72	0.71	0.72
Decision Tree (DT)	0.83	0.85	0.84
Support Vector Machine (SVM)	0.83	0.87	0.84
K-Nearest Neighbors (KNN)	0.81	0.96	0.84
Gauss Naive Bayes (GNB)	0.70	0.70	0.71
Randon Forest (RF)	0.95	0.93	0.95

A.1 Voting Classifier

Figure 8 presents the confusion matrix generated by the Voting Classifier model, highlighting that, out of a total of 254 instances, 223 were correctly classified as 'non-dropouts,' while 31 were confused as 'dropouts.' Likewise, out of a total of 262 instances, 242 were correctly classified as 'dropouts,' but 20 were confused as 'non-dropouts.

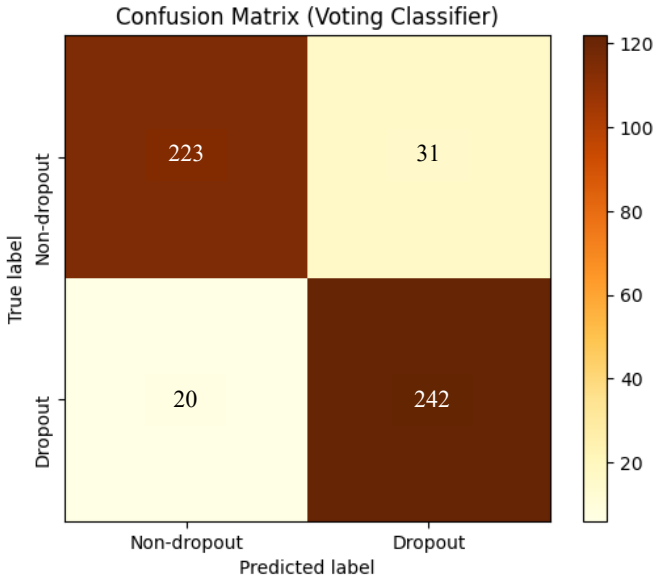

Fig. 8 Voting Classifier Model Confusion Matrix

A.2 Gradient Boosting Classifier

Figure 9 shows the confusion matrix generated by the Gradient Boosting model. In this case, 244 instances were correctly classified with the label 'non-dropouts,' but 10 were

misclassified as 'dropouts.' On the other hand, 228 instances were correctly classified with the label 'dropouts,' while 34 were misclassified as 'non-dropouts.

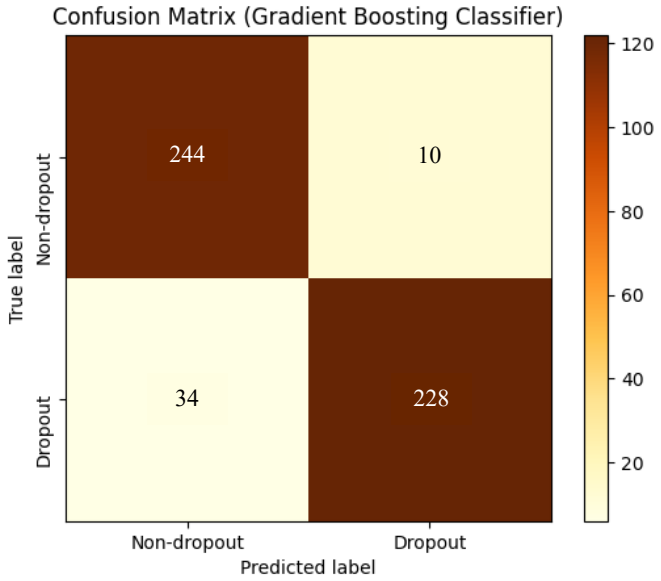

Fig. 9 Gradient Boosting Model Confusion Matrix

A.3 eXtreme Gradient Boosting (XGBoost)

Figure 10 displays the confusion matrix generated by the XGBoost model. Noteworthy among the values are 11 instances labeled as 'non-dropouts' that were confused as 'dropouts,' and 34 instances labeled as 'dropouts' that were confused as 'non-dropouts.

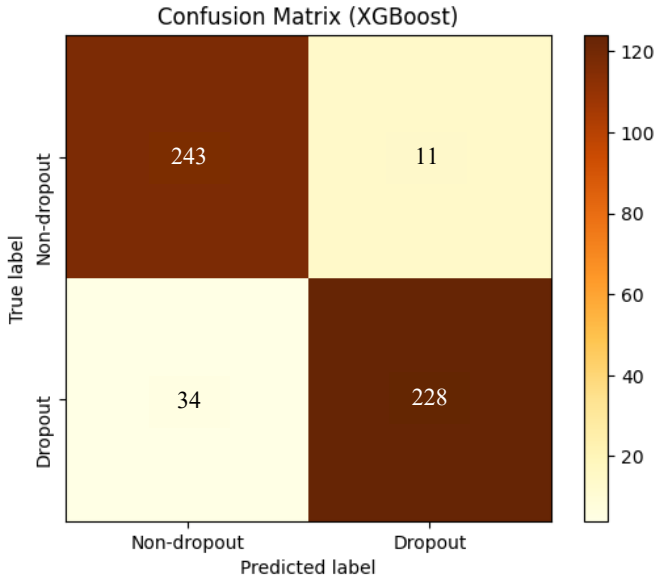

Fig. 10 eXtreme Gradient Boosting Model Confusion Matrix

A.4 LightGBM Classifier

Figure 11 shows the confusion matrix generated by LightGBM. This algorithm stands out among the three previous algorithms for its higher classification precision, correctly classifying 247 instances as 'no dropouts' with their respective labels, and only 7 instances were misclassified. Similarly, 243 instances labeled as 'dropouts' were correctly classified, while 19 instances were misclassified.

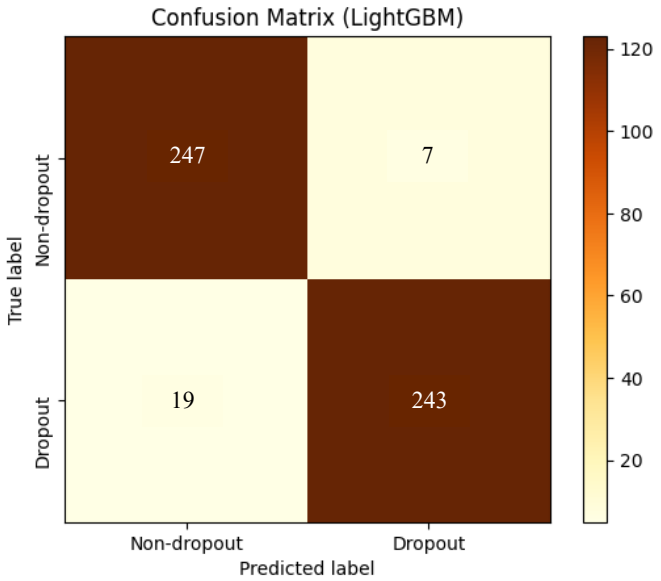

Fig. 11 LightGBM Model Confusion Matrix

A.5 Performance of Classification models

In Table 13, the evaluation metrics obtained for the classification models are presented. These metrics offer a comprehensive view of the performance and effectiveness of each model in predicting the risk of undergraduate student attrition. The figure provides a comparative analysis of the performance metrics, enabling a detailed evaluation and comparison between the models in terms of their predictive ability and effectiveness in identifying the risk of student dropout.

TABLE VI
PERFORMANCE OF VOTING CLASSIFIER ALGORITHMS

Algorithm Classifier	accuracy-score	recall-score	f1_Score
Voting Classifier	0.90	0.92	0.90
Gradient Boosting	0.91	0.87	0.91
eXtreme Gradient Boosting (XGBoost)	0.91	0.87	0.91
LightGBM Classifier	0.95	0.97	0.95

After evaluating the different classification models in terms of accuracy, recall, and F1-score, it is observed that the LightGBM model has demonstrated outstanding performance across multiple assessment metrics. With superior accuracy

and a strong balance between recall and F1-Score, it evidences its ability to predict the risk of undergraduate student dropout more reliably compared to the other models evaluated.

A.6 Computational performance

The computational performance of the LightGBM algorithm was evaluated during the training and prediction process. The model was fitted to the data in 0.0914 seconds during training, and the prediction time to infer results on a test dataset was 0.0052 seconds. These times reflect the efficiency of the algorithm in terms of processing speed, demonstrating its ability to train and make predictions quickly and efficiently in this specific context.

B. Discussion

The implementation of a predictive model to identify and reduce the risk of student dropout at the Army Scientific and Technological Institute in Peru has yielded significant results, offering valuable insights into the effectiveness of various classification algorithms.

The presence of class imbalance in the original dataset posed a significant challenge. As observed in the confusion matrices, the initial model without the SMOTE technique tended to favor the majority class, exhibiting high accuracy rates but compromising its ability to identify dropout cases. The application of SMOTE proved crucial in balancing the classes and enhancing the model's capacity to predict student attrition risk, as evidenced by improved accuracy, recall, and F1-Score metrics. Extensive testing was conducted using diverse algorithms, including Random Forest, Voting Classifier, Gradient Boosting, eXtreme Gradient Boosting (XGBoost), and LightGBM. While each model demonstrated remarkable performance, LightGBM stood out with an exceptional accuracy of 95%, recall of 97%, and F1-Score of 95%. This consistency in metrics suggests that LightGBM is particularly effective for predicting attrition risk in this context.

An evaluation of computational efficiency revealed that the LightGBM algorithm is highly efficient, exhibiting remarkable training and prediction times. This efficiency is crucial for practical implementations and the deployment of real-time systems. In comparison to other works reviewed, our model presents significant advantages, underscoring its ability to effectively identify the risk of student dropout. This superiority suggests that our approach could represent a substantial advancement in the field, providing a more effective tool to address challenges associated with student retention.

Additionally, it is crucial to note that the success of our model may depend significantly on the specific features of the Institute. The unique dynamics of the educational environment and the particularities of the students may influence the generalizability of the results to other academic contexts.

VI. CONCLUSIONS

Student dropout at the undergraduate level has a significant impact worldwide. It not only affects the individuals and institutions directly involved but also has national and international repercussions in terms of economic, social, and scientific development. Addressing this problem is fundamental to fostering equal opportunities, driving innovation, and strengthening human capital at the global level.

The contribution of this study is to propose a predictive model that introduces an innovative approach to reduce the risk of undergraduate student dropout at the Army Scientific and Technological Institute. The study presents a comprehensive analysis of the performance of various ranking models in predicting student dropout. After a detailed evaluation of metrics such as accuracy, recall, and F1-score, it was determined that the LightGBM model stands out for its superior ability to predict, achieving an accuracy of 95%, recall of 97%, and F1-score of 95%, compared to other tested models. This choice is based on the balanced results obtained in several experiments.

In this regard, this proposal efficiently contributes to the institution by enabling the early identification of the risk of student dropout and facilitating the implementation of preventive measures to mitigate factors that negatively impact or hinder students' academic retention.

ACKNOWLEDGMENT

The authors extend their gratitude to the Cybersecurity, IoT, and Artificial Intelligence Research Group (GriCIA) of the Army Scientific and Technological Institute (Instituto Científico y Tecnológico del Ejército) and the Directorate of this university for funding the project.

REFERENCES

[1] D. Andrade-Girón *et al.*, «Predicting Student Dropout based on Machine Learning and Deep Learning: A Systematic Review», *ICST Transactions on Scalable Information Systems*, jul. 2023, doi: 10.4108/eetsis.3586.

[2] A. Kumar, K. K. Eldhose, R. Sridharan, y V. V. Panicker, «Students' Academic Performance Prediction using Regression: A Case Study», en 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, jul. 2020, pp. 1-6. doi: 10.1109/ICSCAN49426.2020.9262346.

[3] J. K. Hoyos Osorio y G. Daza Santacoloma, «Predictive Model to Identify College Students with High Dropout Rates», *Revista Electrónica de Investigación Educativa*, vol. 25, pp. 1-10, may 2023, doi: 10.24320/redie.2023.25.e13.5398.

[4] S. Jayaprakash, S. Krishnan, y V. Jaiganesh, «Predicting Students Academic Performance using an Improved Random Forest Classifier», en *2020 International Conference on Emerging Smart Computing and Informatics (ESCI)*, IEEE, mar. 2020, pp. 238-243. doi: 10.1109/ESCI48226.2020.9167547.

[5] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, y M. Hernandez, «Perspectives to Predict Dropout in University Students with Machine Learning», en *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, IEEE, jul. 2018, pp. 1-6. doi: 10.1109/IWOBI.2018.8464191.

[6] M. R. Rimadana, S. S. Kusumawardani, P. I. Santosa, y M. S. F. Erwianda, «Predicting Student Academic Performance using Machine Learning and Time Management Skill Data», en *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, dic. 2019, pp. 511-515. doi: 10.1109/ISRITI48646.2019.9034585.

[7] M. Baharin, W. R. Ismail, R. R. Ahmad, y N. Majid, «Factors affecting students' academic performance using Analytic Hierarchy Process (AHP)», en *2015 International Conference on Research and Education in Mathematics (ICREM7)*, IEEE, ago. 2015, pp. 169-173. doi: 10.1109/ICREM.2015.7357047.

[8] [8] P. M. da Silva, M. N. C. A. Lima, W. L. Soares, I. R. R. Silva, R. A. de A. Fagundes, y F. F. de Souza, «Ensemble Regression Models Applied to Dropout in Higher Education», en *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, IEEE, oct. 2019, pp. 120-125. doi: 10.1109/BRACIS.2019.00030.

[9] R. Z. Pek, S. T. Ozyer, T. Elhage, T. Ozyer, y R. Alhajj, «The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure», *IEEE Access*, vol. 11, pp. 1224-1243, 2023, doi: 10.1109/ACCESS.2022.3232984.

[10] A. A. Mubarak, H. Cao, y W. Zhang, «Prediction of students' early dropout based on their interaction logs in online learning environment», *Interactive Learning Environments*, vol. 30, n.º 8, pp. 1414-1433, jul. 2022, doi: 10.1080/10494820.2020.1727529.

[11] B. Díaz *et al.*, «Deserción de estudiantes, factores asociados con árboles de decisión: caso Escuela de Postgrado 1de una Universidad pública en Perú», 2022.

[12] O. Nnamdi Iwegbuna *et al.*, «Predicting Students Academic Performance Using Supervised Learning Arduino Aided Design And Programming Of Home Automation System View project Predicting Students Academic Performance Using Supervised Learning», 2022. [En línea]. Disponible en: www.ijrpr.com

[13] Y. Zheng, Z. Gao, Y. Wang, y Q. Fu, «MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series», *IEEE Access*, vol. 8, pp. 225324-225335, 2020, doi: 10.1109/ACCESS.2020.3045157.

[14] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, y F. Sanchez-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.

[15] L. H. Schultz, J. J. Connolly, S. M. Garrison, M. M. Leveille, y J. J. Jackson, «Vocational interests across 20 years of adulthood: Stability, change, and the role of work experiences», *J Res Pers*, vol. 71, pp. 46-56, dic. 2017, doi: 10.1016/j.jrp.2017.08.010.

[16] R. Asif, A. Merceron, S. A. Ali, y N. G. Haider, «Analyzing undergraduate students' performance using educational data mining», *Comput Educ*, vol. 113, pp. 177-194, oct. 2017, doi: 10.1016/j.compedu.2017.05.007.

[17] F. A. Bello, J. Kohler, K. Hinrechsen, V. Araya, L. Hidalgo, y J. L. Jara, «Using machine learning methods to identify significant variables for the prediction of first-year Informatics Engineering students dropout», en *2020 39th International Conference of the Chilean Computer Science Society (SCCC)*, IEEE, nov. 2020, pp. 1-5. doi: 10.1109/SCCC51225.2020.9281280.

[18] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, y G. Van Erven, «Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil», *J Bus Res*, vol. 94, pp. 335-343, ene. 2019, doi: 10.1016/j.jbusres.2018.02.012.

[19] S. M. Merchan Rubiano y J. A. Duarte Garcia, «Formulation of a predictive model for academic performance based on students' academic and demographic data», en *2015 IEEE Frontiers in Education Conference (FIE)*, IEEE, oct. 2015, pp. 1-7. doi: 10.1109/FIE.2015.7344047.

[20] A. Perez, E. E. Grandon, M. Caniupan, y G. Vargas, «Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees», en *2018 37th International Conference of the Chilean Computer Science Society (SCCC)*, IEEE, nov. 2018, pp. 1-8. doi: 10.1109/SCCC.2018.8705262.

[21] G. Kostopoulos, S. Kotsiantis, O. Ragos, y T. N. Grapsa, «Early dropout prediction in distance higher education using active learning», en *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, IEEE, ago. 2017, pp. 1-6. doi: 10.1109/IISA.2017.8316424.

[22] A. Dutt, M. A. Ismail, y T. Herawan, «A Systematic Review on Educational Data Mining», *IEEE Access*, vol. 5, pp. 15991-16005, 2017, doi: 10.1109/ACCESS.2017.2654247.

[23] D. Heredia, Y. Amaya, y E. Barrientos, «Student Dropout Predictive Model Using Data Mining Techniques», *IEEE Latin America Transactions*, vol. 13, n.º 9, pp. 3127-3134, sep. 2015, doi: 10.1109/TLA.2015.7350068.

[24] A. G. Costa, J. C. B. Mattos, T. T. Primo, C. Cechinel, y R. Munoz, «Model for Prediction of Student Dropout in a Computer Science Course», en *2021 XVI Latin American Conference on Learning Technologies (LACLO)*, IEEE, oct. 2021, pp. 137-143. doi: 10.1109/LACLO54177.2021.00020.

[25] H. Liu y W. Zhang, «A Hybrid Deep Learning Model for MOOCs Dropout Pediction», en *2022 4th International Conference on Computer Science and Technologies in Education (CSTE)*, IEEE, may 2022, pp. 178-183. doi: 10.1109/CSTE55932.2022.00039.

[26] E. C. Medina, C. B. Chunga, J. Armas-Aguirre, y E. E. Grandon, «Predictive model to reduce the dropout rate of university students in Perú: Bayesian Networks vs. Decision Trees», en *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, jun. 2020, pp. 1-7. doi: 10.23919/CISTI49556.2020.9141095.

[27] O. Jiménez, A. Jesús, y L. Wong, «Model for the Prediction of Dropout in Higher Education in Peru applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network and Support Vector Machine», en *2023 33rd Conference of Open Innovations Association (FRUCT)*, IEEE, may 2023, pp. 116-124. doi: 10.23919/FRUCT58615.2023.10143068.

[28] A. Kothare, S. Chaube, Y. Moharir, G. Bajodia, y S. Dongre, «SynGen: Synthetic Data Generation», en *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)*, IEEE, nov. 2021, pp. 1-4. doi: 10.1109/ICCICA52458.2021.9697232.

[29] D. Cullen, J. Halladay, N. Briner, R. Basnet, J. Bergen, y T. Doleck, «Evaluation of Synthetic Data Generation Techniques in the Domain of Anonymous Traffic Classification», *IEEE Access*, vol. 10, pp. 129612-129625, 2022, doi: 10.1109/ACCESS.2022.3228507.

[30] S. Kumari, D. Kumar, y M. Mittal, «An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier», *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40-46, jun. 2021, doi: 10.1016/j.ijcce.2021.01.001.

[31] R. Punmiya y S. Choe, «Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing», *IEEE Trans Smart Grid*, vol. 10, n.º 2, pp. 2326-2329, mar. 2019, doi: 10.1109/TSG.2019.2892595.

[32] T. Chen y C. Guestrin, «XGBoost», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.

[33] N. Ghatasheh, I. Altaharwa, y K. Aldebei, «Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction», *IEEE Access*, vol. 10, pp. 84365-84383, 2022, doi: 10.1109/ACCESS.2022.3196905.

[34] K. Zheng, L. Wang, y Z.-H. You, «CGMDA: An Approach to Predict and Validate MicroRNA-Disease Associations by Utilizing Chaos Game Representation and LightGBM», *IEEE Access*, vol. 7, pp. 133314-133323, 2019, doi: 10.1109/ACCESS.2019.2940470.

[35] L. Li, Y. Lin, D. Yu, Z. Liu, Y. Gao, y J. Qiao, «A Multi-Organ Fusion and LightGBM Based Radiomics Algorithm for High-Risk Esophageal Varices Prediction in Cirrhotic Patients», *IEEE Access*, vol. 9, pp. 15041-15052, 2021, doi: 10.1109/ACCESS.2021.3052776.