

Forecasting of Post-Graduate Students' Late Dropout Based on the Optimal Probability Threshold Adjustment Technique for Imbalanced Data

<https://doi.org/10.3991/ijet.v18i04.34825>

Carmen Lili Rodríguez Velasco^{1,2,3}, Eduardo García Villena^{1,4}(✉),
Julién Brito Ballester^{1,2}, Frigidiano Álvaro Durántez Prados^{1,2,3},
Eduardo Silva Alvarado^{1,4}, Jorge Crespo Álvarez^{1,4}

¹Universidad Europea del Atlántico (UNEATLANTICO), Santander, España

²Universidad Internacional Iberoamericana (UNINI-MX), Campeche, México

³Universidade Internacional do Cuanza (UNIC), Kuito, Angola

⁴Universidad Internacional Iberoamericana (UNIB-PR, USA), Puerto Rico
eduardo.garcia@uneatlantico.es

Abstract—Overcoming the predominant analogical models in face-to-face education takes on a special connotation within the e-learning field. The present research contributed in reducing this gap through the development of a predictive model regarding the dropping out of online graduate studies from two universities in the Ibero-American region, using machine learning tools for decision making. In this sense, unlike what happens in a face-to-face approach, the significant variables were identified only with the academic setting in general, and timeliness in particular, excluding the socio-demographic aspects of a student. In line with the Institution's strategy, priority was given to sensitivity or recall, and to adopting the seldom used but effective technique of optimal probability threshold adjustment as opposed to other traditional techniques for processing unbalanced data. In this context, the classifier optimizations were: Logistic Regression, Random Forests and Neural Networks, together with different techniques, attributes, and resampling algorithms (SMOTE, SMOTE SVM, ADASYN and Hyperparameters), provided thresholds between 0.454 and 0.669, sufficiently valid to reach a recall value of 0.75 for the Neural Network classifier with SMOTE_SVM, followed by Logistic Regression with SMOTE_SVM (0.67), and Random Forests with Hyperparameters (0.6). Likewise, with an optimal threshold of 0.427, the robustness of Random Forests for unbalanced classes was demonstrated by achieving metrics very similar to those obtained by consensus of the three previous models (threshold = 0.463). Lastly, this research paper will hopefully contribute in boosting the application of this simple but powerful technique, which is highly underestimated with respect to data resampling techniques for unbalanced classes.

Keywords—optimal likelihood threshold, imbalanced data, student dropout prediction, resample techniques, distance learning courses

1 Introduction

1.1 Dropout from distance university studies

At present, the development of e-learning is a major challenge for educational institutions globally [1]. The lack of knowledge of the methodology involved in e-learning and the digital divide represent two major challenges in the educational field since not all students feel comfortable with virtual procedures, nor do they have equal opportunities in access to technologies [2].

The digital transformation in the university environment is characterized by a progressive increase in the number of students starting online programs. However, in parallel, this upward growth has been accompanied by high dropout and dissatisfaction rates among students [3] that, in some cases, have reached 50% [4] compared to studies on a face-to-face campus [5], [6].

These shortcomings are even more pronounced in massive open online courses (MOOCs), usually free of charge, and which can reach dropout rates above 80% [7].

This fact implies that, nowadays, despite the growing popularity of *online* programs, “retaining students has become a problematic issue” [8], and therefore defining tools for dropout prediction has become essential [9].

Traditionally, graduation and dropout rates in university studies have been considered as a measure of effectiveness, affecting the reputation of educational institutions [10], [11]. In many cases, there is international consensus in considering these indicators as an index of quality [12], which can condition grants and funds from governments, for example, in the USA [12], [13].

The possible causes of dropping out of university studies are complex in nature and involve psychological, social, economic, organizational, and interaction aspects between the environment and the student [14], so sometimes prediction models tend to consider a subset of attributes, which provide a biased idea of the real factors involved in dropout cases [15].

According to its duration, dropping out of university studies can be temporary (*stopout*) or definitive (*dropout*). The boundaries between the two are blurred since sometimes the interruption of studies takes place against a background of commitment to return or reintegration, which can be affected by different circumstances (personal, family, economic...), resulting in a definitive dropout. Conversely, although less likely, an abandonment that is presented as definitive could become temporary if the student later resumes his or her university studies.

Another form of classification is that which refers to individual behavior in which desertion can occur either for academic reasons or voluntarily. In the first case, the student drops out because he/she has not met the academic requirements of the program; on the other hand, voluntary desertion responds to a deliberate act on the part of the student.

A third form of classification refers to the moment at which dropout occurs, distinguishing between initial, early, and late dropout [16]. Some authors also include early dropout to refer to students who, having been accepted by the university, do not enroll [17].

Table 1 summarizes the classification of the forms of university dropout.

Table 1. Classification of the forms of university dropout

| Ranking | Attribute or Characteristic |
|---------------------|-----------------------------|
| Duration | Temporary (stopout) |
| | Definitive (dropout) |
| Individual behavior | Academic reasons |
| | Voluntarily |
| Moment [16], [17] | Initial |
| | Early |
| | Late |
| | Premature |

Rovai [13] defines persistence as “the behavior of continuing action despite the presence of obstacles.” Thus, understood, persistence has a positive connotation since it emphasizes the action of the educational institution as the guarantor of the student’s continuity, while dropout focuses on the individual subject who performs the action of abandoning.

As illustrated in Figure 1, the context of this research article referred to voluntary and late definitive abandonment during the development stage of the Master’s Final Project (MFP), i.e., that which occurs in the last phase of the study program.

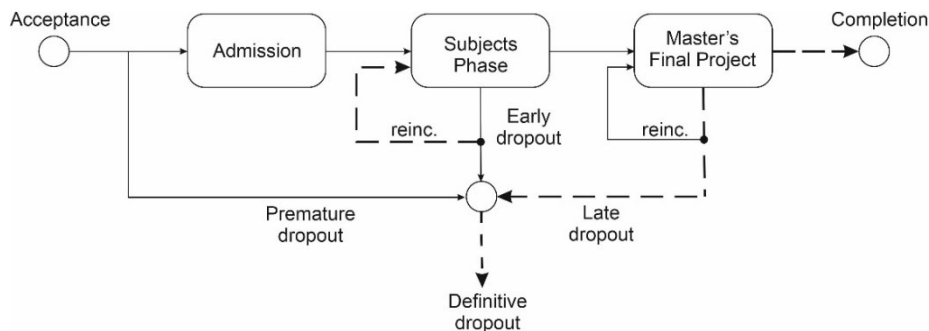


Fig. 1. Life cycle of a student in an educational institution

Note: The stages that, in one way or another, will be taken into account in this research article are shown in dashed lines. Adapted from <http://www.agilemodeling.com/images/models/>

1.2 Statistical analysis vs. Machine learning

Statistical analysis has served for years to study relationships between different variables, validate models, and identify trends in demographic distribution and attributes of the college dropout rate from data extracted from questionnaires, interview transcripts, and other texts [18], [19].

However, these theoretical research methods that, on the one hand, help to identify new predictor variables [18]; present, on the other hand, no few limitations, given their lack of generalization to other institutional contexts and their difficulty of large-scale administration [20].

According to [21], [22], the limitations of using this type of techniques lie in the assumptions required for their application (for example, meeting the criterion of normality of the data) and the difficulties in interpreting and transferring the results of the relationships between variables to the educational community, due to their specificity and the use of technical terms only within the reach of a specialized audience.

In this same context, as it constitutes a reference for a large number of authors [14], it is necessary to mention the model of Vincent Tinto [23]. This model is important for two reasons: firstly, because of the analogy it establishes between the characteristics, expectations, and individual motivations of suicidal behavior and dropout; and, secondly, because of the cost-benefit analysis it performs between the decision to drop out and the decision to persist in studies.

Precisely, interpretability is fundamental in this area since the problem is not so much predicting as preventing dropout. If the characteristics of students and the drivers of learning cannot be explained, effective prevention strategies cannot be prescribed. Without knowing the risk factors, educators cannot tailor their programs or intervene in a personalized way [24].

The diversity of courses, instructional designs, and online platforms, limit the application of conventional statistics to specific learning platforms; therefore, it is essential to create predictive and flexible models that can be adapted to different learning environments [25].

Machine learning (ML), on the other hand, is a more recent field that emphasizes predictive models rather than making inferences about a population from a sample. This represents a number of advantages over conventional forms of statistical analysis [26]. However, the boundaries between statistics and machine learning are very blurred and the topics tend to overlap more and more [27].

Some examples that are topical in Machine Learning, but had their origin in statistics, are correlational analysis [28], [29], logistic, linear and multivariate regression [30], [31], and analysis of variance [32], among others.

As will be seen below, classifiers and, in particular, machine learning, although they require the fulfillment of some conditions, do not need to start from assumptions about the structure of the data, which allows the elaboration of complex nonlinear models with a higher degree of interpretability than statistical models [21].

On the other hand, they analyze the information routinely collected in the databases of educational institutions, complementing the traditional method based on surveys and interviews. In this sense, the institutional repository as a source of data has been gaining more and more prominence in studies on dropout, which has been favored by technological development and the cheapening of information storage devices, which has led public and private organizations to significantly increase the amount of digital information recorded on their users or customers [28].

According to Andrade [33], distance education generates a large amount of data that can serve as raw material for research due to the high level of digital mediation, “[...] a large part of this data has not been analyzed, which constitutes an important gap for conducting research.” It is, therefore, “an unprecedented opportunity for data analytics and *machine learning* (ML) to advance the state of the art” [34].

In this way, both research techniques—the traditional one, based on interviews and surveys, and data analytics, based on the institutional repository—can complement and help each other and, thus, lead to new theories or to an improvement of the existing ones [35].

1.3 Background of educational data mining applied to distance learning dropouts

The increase in the amount of data, technological advances, and the development of analysis tools have meant that in recent years the use of analytical methods and, in particular, data mining [36], [37], has experienced exponential growth as a means of transforming data into information in areas such as marketing, security, and the financial and health sectors, among others.

The challenge is to “develop systems that provide, not only early and accurate alerts but also a justification and explanation of the reasoning behind the decision,” making decision-making understandable to people [24].

In reference to the field of education, mining large datasets has been applied to both face-to-face and *online* teaching to discover unnoticed patterns of behavior in students, extract findings, and find unanticipated relationships between attributes [38].

So much so that Learning Analytics (LA) has become one of the main emerging fields of research for quality improvement in education, introducing analytical methods such as artificial intelligence [39], and which seek to identify valid, useful, and novel patterns of behavior.

It is in this context that machine learning, a new paradigm integral to data mining techniques, such as statistics and artificial intelligence come into play [40]. Among other functions, it allows the development of a predictive model with a large amount of data, which can result in a numerical value (regression) or label a category of data (classification).

Depending on the type of output and treatment approach, machine learning can be presented with examples of inputs and observed outputs (labels), where the objective is that the model trains with this data set and learns to define a general rule that assigns the appropriate output label to a new value [41]. This type of classification, called supervised learning, is the one addressed in this research article.

The literature consulted describes other educational objectives that use these same techniques for classification and that refer, for example, to diagnosing a given teaching strategy, evaluating the quality of a teaching material or discovering students' personal preferences, among others [37]. Consequently, it is urgent that universities adopt these techniques wherever possible [42].

Starting in 2014, the number of papers related to college dropout grew significantly [43]. Until then, most of the research using machine learning techniques had focused on predicting students' grades or their persistence in (mostly *online*) courses [21]. In this context, a research paper conducted in 2019 by teachers at Roma Tre University, on the prediction of dropout in university studies, concluded that the most used classifiers were in this order: Decision Tree (DT), Bayesian Classifiers (NBC), Neural Networks (NN), Logistic Regression (LR), Support Vector Machines (SVM), Miscellanea Algorithms, and K-Nearest Neighbors (KNN) [43].

Table 2 shows some references that use machine learning methodologies to predict college dropout.

Table 2. Research that has made inroads into university dropout using machine learning techniques

| Classifier | References |
|---------------------------------------|-----------------|
| Random Forest | [44]–[48] |
| Bayesian Classification (Naive Bayes) | [49], [50] |
| Neural Network | [51]–[54] |
| Logistic Regression | [51], [55]–[58] |

Note: Own elaboration based on [43], [59].

1.4 The problem of imbalanced classes

Information takes on an exponential character when it comes to studies in a virtual learning environment. In this context, the recording of information is not only referred to the demographic, economic, social, and academic determinants of students (*Educational Data Mining*, EDM), which can be used to find relationships between variables [60], but the Institution also has a massive accumulation of data related to various academic and administrative processes.

In this context, it is common to find problems of classification of the variable to be predicted, where there is a class described as majority or negative, which agglutinates a large proportion of the data, and another minority or positive class, scarcely represented in terms of information, and which usually constitutes the class of interest.

Working with imbalanced data in relation to machine learning is a problem of growing interest [61] since in these cases algorithms tend to classify all observations as instances of the majority class [62], which results in low recall for the class of interest [63] and can lead to errors and poor generalization of the model’s behavior.

The most favorable solution is to extend data collection; however, this is not always possible, so one or more of the following combined techniques must be used (Table 3):

- Resampling: a uniform distribution between classes is achieved by altering the data distribution of the model. This technique has the disadvantage that it can introduce examples of the minority class in the majority class and cause, in practice, problems of overfitting or underfitting, which could invalidate the model.
- Hyperparameter penalty: a higher weight is given to the minority class to the detriment of the majority class. It does not alter the data distribution of the model.
- Optimal probability threshold setting: a fair probability of occurrence is assigned. It does not alter the data distribution of the model.

Table 3. Some imbalanced data processing techniques and associated references

| Approach | Technique | | Method | References |
|--|--|-------------------------------|---|-----------------|
| Resampled Models (alter data distribution) | Under-sampling | Random Under-Sampling (RUS-I) | Random elimination of instances of the majority class. No replacement. | [64]–[66] |
| | | Tomek Links | Only instances of the majority class that are redundant or very close to instances of the minority class are eliminated. | [67]–[69] |
| | Over-sampling | SMOTE | Based on the K-Nearest-Neighbors classifier. Synthesis of new instances by random selection of nearest neighbors and interpolation within the minority class. | [64], [70]–[73] |
| | | SMOTE SVM | Synthesis of new instances within the minority class, based on the creation of a classification hyperplane of two or more categories. | [74]–[76] |
| | | ADASYN | Synthesis of new instances within the minority class, based on the density distribution of the data. | [77], [78] |
| | Hybrid methods | SMOTE-Tomek Links | Tomek’s links apply to oversampled minority class samples made by SMOTE. | [79], [80] |
| Base Models (do not alter the data distribution) | Penalty for hyperparameters | | Search for hyperparameters that improve the efficiency of the model by favoring the minority class. | [81]–[83] |
| | Setting the optimal prediction probability threshold | | Search for an optimal prediction probability threshold, in an imbalanced data scenario, where the default threshold of .5 is not adequate. | [63], [84]–[87] |

As can be seen, most of the literature found in this research refers to the use of resampling techniques to deal with imbalanced data. Unlike those, the adjustment of an optimal prediction probability threshold has hardly been addressed by researchers, despite being equally a valid and efficient alternative [63], [86]. In this sense, no study referring to college dropout has been found that has been addressed under this perspective.

1.5 Research design

The objective of this research article was to contrast the benefits of the optimal probability threshold adjustment technique, with other non-balanced data processing techniques, in its application to the prediction of late dropout from distance learning graduate studies.

Thus, once the problem and objective were stated, the research question was as follows:



- Is it possible to obtain equivalent metrics when comparing optimized base models with a prediction probability threshold, with other complex models, consensual from different classifiers?

An attempt was also made to answer the following research sub-questions:

- What variables have a determining influence on late dropout from university studies?
- Can we contribute to the development of early prevention measures aimed at avoiding student dropout during the final phase of their postgraduate distance learning studies?
- Is it appropriate to use the default probability threshold for imbalanced data?
- Is it possible to use the dropout prediction probability threshold adjustment technique in conjunction with other resampling techniques such as SMOTE, for example?

The guidelines for this research are shown in Table 4.

Table 4. Research design

| | |
|--|--|
| Unit of analysis: | 3 base models—Logistic Regression, Random Forest, and Neural Network-, optimized with varying prediction probability thresholds, and in combination with techniques for the treatment of non-balanced data. Application to the late dropout of postgraduate university studies at Universidad Europea del Atlántico, UNEATLANTICO (Spain) and UNINI-MX (Mexico) during the 2013–2019 period. |
| Dependent variable: | Academic Status (MFP stage) Operational definition of the dependent variable |
| Values of the dependent variable: |  Majority class: Finishes Minority class or class of interest: Dropout |
| Independent variables of the student: | Demographic and personal variables General academic variables Academic variables related to the MFP Academic variable of completion/dropout |
| Observation unit: | What is the source of the data on late college completion or dropout? <div style="text-align: center;">  </div> Institutional repository database consisting of 12,370 students, belonging to 34 different graduate programs. |

Note: Adapted from [88], [89].

Thus, the article was structured in different parts. In the Introduction, a theoretical approach was made to the abandonment of distance university studies, the role of traditional statistics, and how it complements artificial intelligence in its analysis. In this sense, the background involving data mining and machine learning served as a preamble to the processing techniques for imbalanced data and classifiers, whose objective was to provide a set of metrics and graphs to agree on an optimal prediction probability threshold, which was contrasted with the base models to observe similarities and differences. Finally, the discussion section was followed by the conclusions, including the results and their interpretation.

Finally, it is worth mentioning that during the course of this research, studies with notable shortcomings on the issue of school dropout were found. In this context, [90] found studies that either obviated class imbalance, relied on metrics such as accuracy, or drew their conclusions based on one of the worst rated classifiers for imbalanced data scenarios, such as Logistic Regression.

Despite the fact that, at the present time and at this level, no references have been found on the adjustment of the optimal threshold of probability of predicting late dropout from university studies, we have tried as far as possible to contrast and handle the information with the rigor required by research of this type.

2 Materials and methods

2.1 Introduction

In order to meet the objectives and answer the research questions, we evaluated the ability of three classifiers (Logistic Regression, Random Forest, and Neural Network) to predict the correct class—dropout or completion—from a set of student characteristics, in combination with different resampling techniques (RUS, SMOTE, SMOTE SVM, ADASYN), hyperparameter penalization, and adjustment of the optimal dropout prediction probability threshold.

Broadly speaking, after data processing, the following steps consisted of training the model using the stratified cross-validation technique with data resampling techniques, finding the prediction probabilities for minority classes, ranking these probabilities using an iterative process based on a range of thresholds from 0 to 1 with one step $\epsilon = .001$, determination of the (optimal) threshold value that maximizes the f1-Score metric [85], validation of the model with the obtained threshold value, determination of metrics, establishment of a consensus among the three best models that represented the decrease in the false negative rate, and finally, determination of a consensus threshold to contrast similarities and differences with the fitted base models.

The study responded, therefore, to a descriptive and relational methodology, with a quantitative, non-experimental, transactional approach, because no hypotheses were proposed and no variables were manipulated, but “[...] data were measured, evaluated or collected on various aspects, dimensions or components of the phenomenon to be investigated [in their natural work environment and in a single time]” [91], [92].

2.2 Population and sample

The study population was constituted in December 2020, from a total of 12,370 students enrolled jointly from the European Atlantic University of Santander (Spain) and the International Ibero-American University (UNINI-MX), belonging to 34 different postgraduate programs taught in online mode between the years 2013 to 2019 and in the beginning phase of the MFP.

After eliminating duplicate records and those corresponding to unavailable data (n/a), the final sample consisted of 10,934 students.

For this sample, a total of 12 independent characteristics were selected from the institutional repository, in addition to the binary dependent variable corresponding to the situation of late dropout at the stage of the MFP or completion of university studies by the students.

2.3 Matrix of characteristic variables

Based on their intrinsic nature and interaction with the environment, the characteristic variables used in this research article were classified as shown in Table 5.

Table 5. Matrix of variables characteristics

| Type of Variable | Concept | Features | Operational Definition |
|------------------|--|---------------------------------------|---|
| Independent | Demographic and personal variables | Age | Numerical variable indicating the age (in years) of entry into the program |
| | | Genre | Binary variable that indicates whether the student is male or female |
| | | Origin | Qualitative variable indicating the geographic location of the student |
| | | Employment status | Binary variable indicating whether the person is active in the labor market or unemployed at the time of enrollment |
| | | Level of education | Binary variable indicating the highest level of studies completed when enrolled in the program (undergraduate or postgraduate) ¹ |
| | General academic variables | Academic department | Qualitative variable indicating the Academic Department to which the program belongs |
| | | Qualification | Binary variable indicating whether the person is enrolled in UNEATLANTICO or UNINI-MX |
| | | Extensions | Quantitative variable that indicates the number of academic and/or contracted extensions for the extension of studies |
| | | Reincorporation | Binary variable indicating whether or not the student, after a period of leave, has rejoined his or her program of studies during the initial stage |
| | | Duration block of subjects | Numerical variable indicating the time elapsed (in months) from the date of enrollment in the program to the start of the MFP |
| | Academic variables related to the MFP | MFP duration | Numerical variable that indicates the time elapsed (in months) between the start date and the end or dropout date of the MFP |
| | | Stability in the direction of the MFP | Binary variable that indicates whether there has been only one or more than one Director of the MFP during the development stage of the MFP |
| Dependent | Academic variable of completion/ dropout | Academic status | Binary variable that indicates whether the student has finished or definitively abandoned his/her university studies during the stage of development of the MFP |

Note: ¹Master’s or higher education is required for the variable to take the value “Postgraduate degree”.

2.4 Data preparation

During data preparation, a typical situation of imbalanced data was observed, with the majority class accounting for 74.8% of the sample and the minority class for the remaining 25.2% (Figure 2).

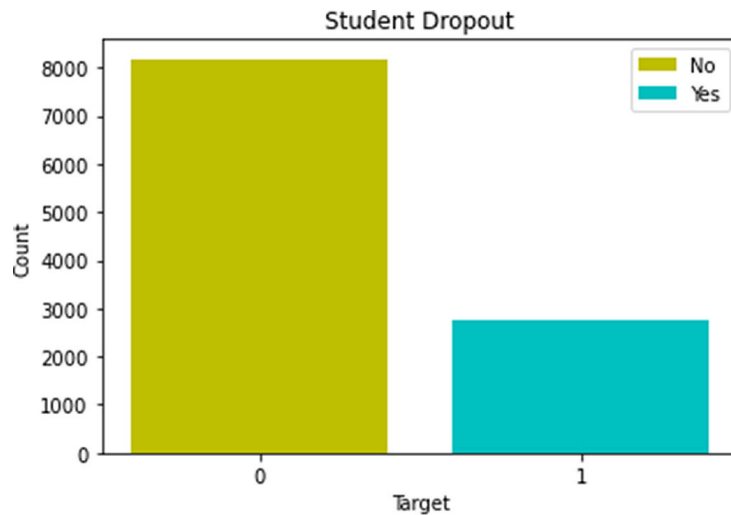


Fig. 2. Ratio between the majority and minority classes in the sample

Subsequently, after converting the categorical qualitative variables “Origin” and “Academic Department” (dummy variables) to numerical format and deleting the original ones, we proceeded to arbitrarily eliminate one of the respective dummy variables, in order to avoid the multicollinearity trap. After this process, the total number of independent variables grouped in the characteristic’s matrix “X” was 19, while the only dependent variable “Academic Status” constituted the vector “y” or target.

2.5 Division of data into training and testing

In this phase, the data from the feature matrix were divided into training (70%) and testing (30%).

Data training was performed using the stratified cross-validation technique. This process was useful to minimize the possibilities of overfitting. For the purposes of comparability, these data sets were kept unchanged in all the scenarios described below.

2.6 Selection of significant model variables

In order to reduce the dimensionality of the model without losing information, the most significant variables that were candidates to be part of the model were categorized according to their importance, using the training data of the Recursive Feature Elimination Cross-Validated (RFECV) algorithm, included in the Scikit-learn library of Python v3.10.

In this way, a total of 10 independent variables were finally determined as the most significant, which, together with the objective or target variable, were the ones that formed part of the research.

2.7 Model training

This stage consisted of several previous steps:

Selection of processing techniques for imbalanced data. The techniques employed in each scenario for the treatment of imbalanced data described in this research article were carried out by implementing the Imbalanced-learn library of Python v3.10. They consisted of one or more combinations of the following:

1. Increasing the sample size of the minority class (Oversampling).
2. Decrease in the sample size of the majority class (Undersampling).
3. Hybrid methods (combination of Oversampling and Undersampling techniques).
4. Penalty for hyperparameters adjusted by the user.
5. Synthetic data generation using SMOTE, SMOTE_SVM, and ADASYN.
6. Determination of an optimal prediction probability threshold.

It is important to mention that these techniques were applied only to the training data set since it is a common mistake to do so on the testing data as well, prior to splitting the data, thus resulting in an over-fitted and poorly generalizable model [78].

Classifier selection. As mentioned above, the classifiers described in this research article were Logistic Regression, the Random Forest algorithm, and the Neural Network of two or more hidden layers (Deep Learning).

All of them are based on minimizing-through the gradient descent technique, Newton-Raphson, etc.—a loss function during training, which calculates the cross-entropy loss between the current and predicted observations [93], in order to obtain the parameters or weights that best fit the model.

In this context, binary logistic regression predicts the probability of occurrence of an event or class, conditional on a set of “n” independent variables, according to equation 1:

$$P_i(X) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (1)$$

where X_i is each of the characteristics of the model and β_i the weight of each of them.

As mentioned above, this probability is classified according to a threshold (δ) in two categories of the response or target variable: zero ($P(X_i) < \delta$) or one ($P(X_i) \geq \delta$).

The Random Forest classifier uses the average of a series of individual decision trees in several subsamples, to obtain the prediction of a new observation once trained. It is a technique that generally improves the result and is widely used to work with imbalanced data [94]. However, an excessive number of trees can lead to overfitting of the model [95].

Finally, the neural network classifier translates into a mathematical algorithm consisting of a series of connected processing units or neurons in which a linear

combination of the weights multiplied by the inputs is performed to subsequently determine the output based on the resulting sum, by means of a continuous, differentiable, and nonlinear activation function (sigmoid, ReLu, hyperbolic tangent...), obtaining values as close as possible to 0 and 1 at the output. Therefore, analogous to logistic regression, the ultimate goal of the neural network is to find the weights and biases that minimize the loss function.

Table 6 shows the configuration of the classifiers used in this research article.

Table 6. Parameters and special configurations of classifiers

| Algorithm | Strategy | Implementation Python 3.10 |
|---------------------|----------------------|---|
| Random Forest | None | class_weight = None, n_estimators= 100, min_samples_split= 2, min_samples_leaf=1, max_features= "sqrt", max_depth= None, criterion = "gini" |
| | Best Hyperparameters | class_weight= "balanced_subsample", n_estimators= 143, min_samples_split= 18, min_samples_leaf=1, max_features= "auto", max_depth= 170, criterion = "gini" |
| Logistic Regression | None | classifier = LogisticRegression (class_weight = None) |
| | Weight Class | classifier = LogisticRegression (class_weight = 'balanced' ...) |
| Neural Network | Dropout1 | Input and first hidden layer: units (10), activation function (ReLu), input_dim=10, dropout =.1 Second hidden layer: units (10), activation function (ReLu), dropout =.1 Output layer: units = 1, activation function (sigmoid) |

Note: Dropout layers randomly set input units to 0 with a t frequency at each step during training time, which helps prevent overfitting.

Optimal probability threshold of prediction. It is a parameter that for each probability value $p(i) \in [0,1]$ assigns a discrete class label [0-end] [1-leave]. As discussed, by default, the prediction probability threshold is 0.5, such that:

$$Y_p = \begin{cases} 0, & p(i) \leq .5 \\ 1, & p(i) > .5 \end{cases} \quad p(i) \in [0,1] \quad (2)$$

As with the processing techniques for imbalanced data, the optimal prediction probability threshold was obtained from the training data set.

Performance indicators. The performance of a classifier is shown in the confusion matrix (Table 7).

Table 7. Confusion matrix for a binary classification model

| | | Predicted Class | |
|--------------|---|---------------------|---------------------|
| | | 0 | 1 |
| Actual Class | 0 | True Negative (tn) | False Positive (fp) |
| | 1 | False Negative (fn) | True Positive (tp) |

As shown in Table 8, the confusion matrix is used to develop the model’s performance evaluation metrics.

Table 8. Performance evaluation metrics for a binary ranking model

| Metrics | Description |
|--|---|
| $Overall\ Accuracy\ rate = \frac{tp + tn}{tp + fp + fn + tn}$ | Overall hit percentage. Not a good indicator for imbalanced data |
| $Individual\ Accuracy\ for\ class\ 0 = \frac{tn}{tn + fn}$ | Individual hit percentage per class. Can be used for imbalanced data |
| $Individual\ Accuracy\ for\ class\ 1 = \frac{fp}{fp + tp}$ | |
| $Sensitivity\ (recall) = \frac{tp}{(tp + fn)}$ | Proportion of positive cases correctly identified by the classifier. Determines when false negative costs are high |
| $Specificity = \frac{tn}{(tn + fp)}$ | Proportion of negative cases correctly identified by the classifier |
| $Precision = \frac{tp}{(tp + fp)}$ | Model quality level. Determines when false positive costs are high |
| $f1\ -\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall}$ | It is used to easily compare measures of precision and sensitivity in a single value. It is very useful for binary classification problems, where the study is focused on the positive class as is the case |
| Receiver Operating Characteristics (ROC) and Area Under Curve (AUC) | ROC is a probability curve that represents on the abscissa axis the <i>fp</i> rate and, on the ordinate axis, the <i>tp</i> rate, for different thresholds. It indicates how much the model is able to distinguish between classes. The area under the AUC curve classifies the performance. The closer AUC is to unity, the better the model distinguishes between classes |
| P-R (precision-recall) Curve | It allows to relate recall and precision. It is interesting that both values are as high as possible; however, the increase of one leads to the decrease of the other |

In this research article, the value of precision, recall, and f1-Score were taken as a reference to determine the generalization capacity of the models since the overall accuracy does not represent an adequate metric for working with imbalanced data [96], [97].

In imbalanced data, the objective is to increase recall without losing precision. However, as will be seen, this is contradictory since a decrease in one leads to an increase in the other and vice versa. In this sense, a choice will have to be made between increasing recall or precision, depending on the circumstances.

In this context, the objective is to establish models with a balanced level of false positives and false negatives, especially in terms of reducing the rate of the latter, optimizing sensitivity (recall) over precision since, as will be seen in the results section, there is a high initial cost associated with false negatives, i.e., predicting that the student will finish when in fact he/she is dropping out of school.

From data training to setting the optimal predictive probability threshold. As illustrated in Figure 3, the training set was divided into $k=3$ groups for each of the classifiers. In each of the groups, $k-1$ parts were taken for training and 1 for testing. The process was repeated a total of k times, rotating the testing set each time. Applying the corresponding events from the testing set to the model resulted in k -matrices (10934×2) of $p(i) \in [0,1]$ probability, with one column for the majority class and one column for the minority class. The minority class probabilities were then averaged and ranked using an iterative method according to a range of probability thresholds (δ) between 0 and 1, with a step $\varepsilon = .001$. These binary values were compared with the observed responses of the target variable y_{train} to find the f1-Score metric, which was stored in a vector of one thousand components from which the maximum value and its threshold were chosen, corresponding to the optimum required for each particular model.

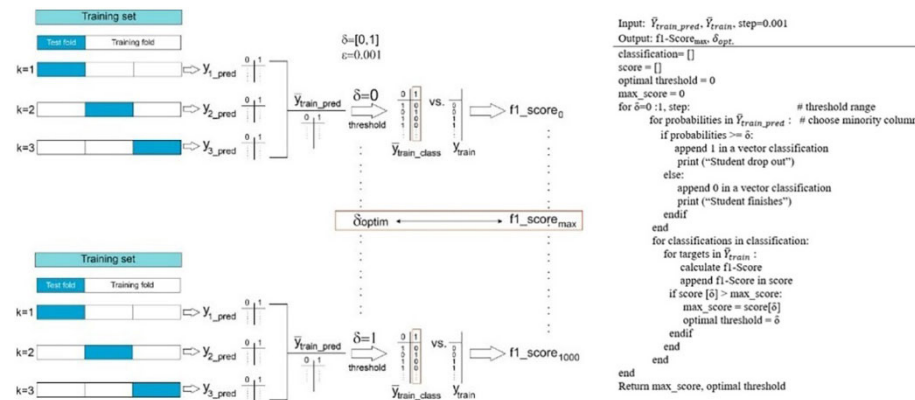


Fig. 3. Illustration (left) and Pseudocode (right) of the training stage of the models resampled by stratified cross-validation and determination of the optimal threshold of probability of predicting dropout in each case

Determination of the consensus optimal threshold of probability of dropout prediction. The consensus optimal threshold for the probability of predicting dropout was determined from the selection of the three models—one for each classifier—that maximized the recall value. In this sense, the arithmetic mean function was used to combine the prediction probabilities, which reduced the error variance in the continuous interval $[0,1]$ and they were classified according to a range of thresholds between 0 and 1 [98], [99]. Finally, the optimal threshold corresponding to the maximum value of the f1-Score metric was found (Figure 4).

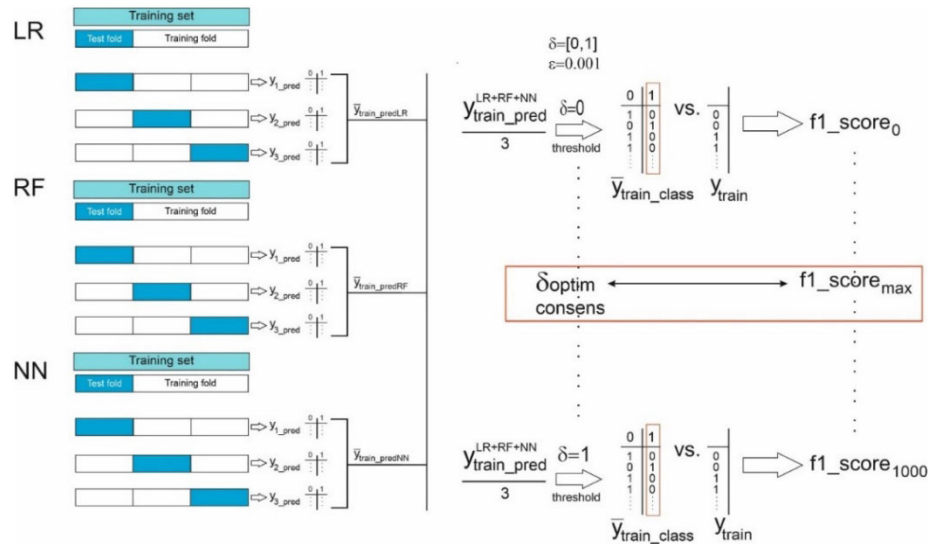


Fig. 4. Determination of the optimal consensus threshold for the probability of predicting late dropout from university studies

2.8 Model validation

Once the model was trained and the optimal probability threshold for prediction in each case was determined, it was introduced into the model and validated using test data. In this way, the confusion matrix and associated metrics were found in each case. In this context, the consensus optimal threshold for the probability of predicting dropout was implemented in the base models and the resulting metrics were compared.

3 Results

3.1 General student characteristics

Table 9 shows the nominal categorical variables and their characteristics used in this research article.

Table 9. Categories and characteristics of nominal independent variables

| Nominal Variables | Attribute | Category | n | Percentage |
|--|-----------|-----------------------|-------|------------|
| Gender | 0 | Male | 4,733 | 43.28% |
| | 1 | Female | 6,201 | 56.71% |
| Origin | 0 | North America | 840 | 7.68% |
| | 1 | Central-South America | 9,296 | 85% |
| | 2 | Africa | 446 | 4% |
| | 3 | Eurasia | 352 | 3.22% |
| Academic Department | 0 | Education and FP | 3,253 | 29.75% |
| | 1 | Company | 2,968 | 27.14% |
| | 2 | Environment | 998 | 9.12% |
| | 3 | Projects | 1,432 | 13% |
| | 4 | Health | 1,490 | 13.62% |
| | 5 | ICTs | 633 | 5.78% |
| | 6 | Tourism | 160 | 1.46% |
| Certification | 1 | UNEAT | 7,723 | 70.63% |
| | 0 | UNINI-MX | 3,211 | 29.36% |
| Reincorporation (course implementation phase) | 1 | Reinstated | 4,726 | 43.22% |
| | 0 | Not reinstated | 6,208 | 56.77% |
| Stability in the Direction of the MFP | 1 | 1 director | 7,683 | 70.26% |
| | 0 | More than 1 director | 3,251 | 29.73% |

The table shows that the proportion of women pursuing graduate degrees at both universities—UNEATLANTICO and UNINI-MX—is thirteen points higher than that of men. Likewise, students from Latin and Central America make up more than three quarters of the sample studied, and the departments of Education and Teacher Training (FP), Business and Health, in that order, have the highest acceptance. As for the University, slightly more than 70% of the population enrolled through UNEATLANTICO, while approximately 30% enrolled through UNINI-MX.

Students who returned to their studies after a period of inactivity accounted for 43.22%, while those who completed the program without leaving constituted 56.77%. Finally, 70% of the students kept the same thesis director throughout, while approximately the remaining 30% changed on at least one occasion for different reasons.

Table 10 shows the ordinal categorical independent variables and their characteristics.

Table 10. Categories and characteristics of the ordinal independent variables

| Ordinal Variables | Attribute | Category | N | Percentage |
|----------------------|-----------|----------------------|--------|------------|
| Age group (years) | 0 | 19–29 | 3,422 | 31.3% |
| | 1 | 30–39 | 4,244 | 38.81% |
| | 2 | 40–49 | 2,287 | 20.91% |
| | 3 | 50–59 | 846 | 7.73% |
| | 4 | 60–69 | 128 | 1.17% |
| | 5 | 70–79 | 7 | .06% |
| Admission profile | 0 | Degree/Dip./Bachelor | 10,679 | 97.66% |
| | 1 | Postgraduate | 255 | 2.33% |
| Employment situation | 0 | Unemployed | 367 | 3.35% |
| | 1 | With employment | 10,567 | 96.64% |

The table shows that practically 60% of the sample under study is between 30 and 49 years of age, most with a degree or diploma and a job.

The quantitative variables were grouped into categories for better understanding (Table 11).

Table 11. Predictor or independent variables of a quantitative nature

| Quantitative Variables | Category | n | Percentage |
|-------------------------------------|----------|-------|------------|
| Extensions | 0 | 2,655 | 24.3% |
| | 1–4 | 7,388 | 67.57% |
| | 5–9 | 828 | 7.57% |
| | 10–14 | 58 | .53% |
| | 15–20 | 5 | .04% |
| Duration block of subjects (months) | 0–19 | 6,479 | 59.25% |
| | 20–39 | 3,916 | 35.81% |
| | 40–59 | 393 | 3.59% |
| | 60–79 | 119 | 1.088% |
| | 80–99 | 22 | .20% |
| | 100–109 | 2 | .018% |
| | 110–129 | 3 | .027% |
| Duration of the MFP (months) | 0–9 | 4,049 | 37.03% |
| | 10–19 | 4,533 | 41.45% |
| | 20–29 | 1,368 | 12.51% |
| | 30–39 | 514 | 4.7% |
| | 40–... | 470 | 4.3% |

The table shows that only 24.3% of the sample under study did not have to request any extension of time to finish. Likewise, 59.25% took between 0 and 19 months to complete the first phase of the program prior to starting the MFP, and only 37.03% took between 0 and 9 months to finish it (or abandon it).

3.2 Correlation matrix of independent variables

Figure 5 shows the correlation matrix of the independent variables.

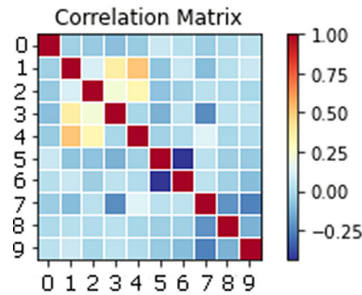


Fig. 5. Correlation matrix of the independent variables

Note: 0. “Certification”; 1. “Extensions”; 2. “Reincorporation”; 3. “Duration_subjects”; 4. “Duration_MFP”; 5. “Origin_Eurasia”; 6. “Origin_Latam_Centr.”; 7. “Department_Education_and_FP”; 8. “Environmt Department”; 9. “Health_Department”.

It can be observed that, in general, there is no significant dependence ($>.5$) between the independent variables, which is very satisfactory.

3.3 Categorization of significant variables

Figure 6 (left) shows a categorization of the ten most significant variables, provided by the Recursive Feature Elimination, Cross-Validated (RFECV) algorithm. Also, the graph shows that the optimal number of variables to choose is 5 (right).

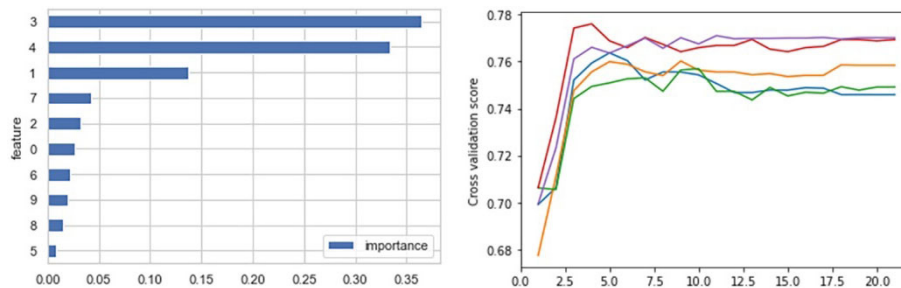


Fig. 6. Categorization of the ten most significant variables for model construction (left) and determination of the optimal number of variables (right)

Note (in order): 3. “Duration_subjects”; 4. “Duration_MFP”; 1. “Extensions”; 7. “Department_Education_and_FP”; 2. “Reincorporation”; 0. “Certification”; 6. “Latam_Central_Origin”; 9. “Health_Department”; 8. “Environment_Department”; 5. “Origin_Eurasia”.

It can be seen how the variables, “Duration_subjects”; “Duration_MFP”; “Extensions”; “Department_Education_and_FP”; “Reincorporation”; and “Certification”; in that order, explain around 90% of the late dropout from university distance learning studies. However, for the purposes of this research article, and in order to lose as little information as possible, the model of ten independent variables was selected.

3.4 Finding the optimal threshold of probability of predicting university dropout

As an example, Table 12 shows the metrics obtained by the base Logistic Regression model for a default prediction probability threshold of 0.5, for both the majority and minority classes.

Table 12. Logistic Regression classifier metrics (baseline model)

| | Precision | Recall | f1-Score | Support |
|-------------------|------------------|---------------|-----------------|----------------|
| Majority class: 0 | .79 | .96 | .86 | 2454 |
| Minority class: 1 | .64 | .24 | .35 | 827 |
| Accuracy | — | — | .77 | 3281 |
| Macro avg | .72 | .60 | .60 | 3281 |
| Weighted avg | .75 | .77 | .73 | 3281 |

In general, although the model seems to respond well to the predictions of the majority class, this is not the case for the minority class. This is a typical case of imbalanced data, with a high precision value in the majority class and a low recall in the minority class.

As can be seen, the overall accuracy value is 0.77. However, it is an inaccurate indicator for imbalanced data, so other more appropriate metrics such as recall or f1-Score will be used for the interpretation of the results [100].

As for the *f1-Score*, it is appropriate when there is a significant problem with false negatives (low *Recall*) and also provides a more reliable assessment of model performance for imbalanced data since, unlike overall accuracy, it takes into account the distribution of the data, although it is more difficult to interpret as it is a harmonic mean between Precision and Recall.

Henceforth, unless otherwise stated, metrics will always refer to the class of interest or minority interest.

Table 13 shows the results of the base classifier metrics after adjustment with the default prediction probability threshold.

Table 13. Metric results for the three base classifiers with the default prediction probability threshold of 0.5

| Classifier | Default Threshold | Overall Accuracy | Precision | Recall | f1-Score | ROC/AUC |
|------------------------------|-------------------|------------------|-----------|--------|----------|---------|
| Logistic Regression baseline | .5 | .77 | .64 | .24 | .35 | .59 |
| Random Forest baseline | .5 | .79 | .60 | .47 | .53 | .68 |
| Neural Network baseline | .5 | .80 | .68 | .42 | .52 | .67 |

From the overall accuracy metric, one might think that the model does a good job in all three cases; however, the low recall values indicate that there is a high false negative rate, so that most of the minority class is not recognized. This is a typical problem for imbalanced classes, where none of the three models could be acceptable.

In relation to the f1-Score metric, only in the base Logistic Regression model does the indicator appear to be significantly lower, probably due to the disproportion between false positives and false negatives in the minority class.

Since the cost associated with a false positive is lower in this case than the cost associated with a false negative, recall should be optimized over precision.

Thus, Table 14 shows the results of the base classifier metrics after adjustment with the corresponding optimal prediction probability threshold.

Table 14. Results of the metrics of the three base classifiers after adjustment with the corresponding optimal probability threshold

| Classifier | Optimal Threshold | Overall Accuracy | Precision | Recall | f1-Score | ROC/AUC |
|------------------------------|-------------------|------------------|-----------|--------|----------|---------|
| Logistic Regression baseline | .243 | .74 | .48 | .65 | .55 | .70 |
| Random Forest baseline | .427 | .78 | .56 | .55 | .55 | .70 |
| Neural Network baseline | .332 | .80 | .59 | .61 | .60 | .81 |

As can be seen, the values of the area under the curve (AUC) in the Logistic Regression and Neural Network classifiers have increased significantly relative to the previous case, although by itself this is not a useful metric to classify performance in the case of imbalanced data, it also needs the f1-Score [86].

In all three cases there is a significant increase in recall (decrease in false negatives) and a decrease in precision (increase in false positives), thus achieving a balance between both metrics, much more evident in the Random Forest and Neural Network classifiers than in Logistic Regression.

Table 15 shows the set of metrics obtained-ordered in decreasing order of recall-, for the case of the resampled and fitted models under optimal probability threshold conditions.

Table 15. Set of metrics for the resampled and adjusted models

| CLASS. | SMOTE | SMOTE_SVM | ADASYN | HYPERP. | OTHER(*) | Thresh (optimal) | tp | tn | fp | fn | Accuracy | Precision | Recall | f1-Score | ROC/AUC |
|--------|-------|-----------|--------|---------|----------|------------------|-----|------|-----|-----|----------|-----------|--------|----------|---------|
| NN | | √ | | | | .512 | 621 | 1821 | 630 | 209 | .74 | .50 | .75 | .60 | .84 |
| | | | √ | | | .604 | 594 | 1886 | 565 | 236 | .76 | .51 | .72 | .60 | .84 |
| | √ | | | | | .585 | 559 | 1994 | 457 | 271 | .78 | .55 | .67 | .61 | .84 |
| LR | | √ | | | | .454 | 551 | 1818 | 636 | 276 | .72 | .46 | .67 | .55 | .70 |
| | √ | | | | | .472 | 531 | 1803 | 651 | 296 | .71 | .45 | .64 | .53 | .68 |
| | | | | | √ | .490 | 523 | 1911 | 543 | 304 | .74 | .49 | .63 | .55 | .70 |
| RF | | | √ | √ | | .571 | 507 | 2001 | 453 | 320 | .76 | .53 | .61 | .57 | .71 |
| | | | | √ | | .555 | 496 | 2077 | 377 | 331 | .78 | .57 | .60 | .58 | .72 |
| LR | | | √ | | | .518 | 491 | 1906 | 548 | 336 | .73 | .47 | .59 | .53 | .68 |
| RF | √ | | | √ | | .591 | 476 | 2084 | 370 | 351 | .78 | .56 | .58 | .57 | .71 |
| | | √ | | √ | | .601 | 457 | 2105 | 349 | 370 | .78 | .57 | .55 | .56 | .70 |
| | | √ | | | | .577 | 440 | 2055 | 399 | 387 | .76 | .52 | .53 | .53 | .68 |
| | √ | | | | | .626 | 410 | 2080 | 374 | 417 | .76 | .52 | .50 | .51 | .67 |
| | | | | | √ | .669 | 411 | 2148 | 306 | 416 | .78 | .57 | .50 | .53 | .68 |
| | | | √ | | | .658 | 393 | 2081 | 373 | 434 | .75 | .51 | .48 | .49 | .66 |

Notes: NN: Neural Network; LR: Logistic Regression; RF: Random Forest. The best false negative reduction models for each of the classifiers are highlighted in shading. (*) Other includes the implementation of a proprietary balancing algorithm (class_weight = ‘balanced’) for Logistic Regression and hybrid resampling (Oversampling/Undersampling) for Random Forest.

The best models that minimize the presence of false negatives are:

- Neural Network with SMOTE SVM
- Logistic Regression with SMOTE SVM
- Random Forest with Hyperparameters

Although the latter is not the one that achieves maximum recall within the Random Forest group, this model was chosen because it has a better overall metric than the one that combines ADASYN and the hyperparameters.

Apart from this, a significant improvement in recall is observed in all cases, produced by incorporating individual hyperparameters, adjustments with probability thresholds, or combinations with resampling techniques in relation to the base model.

In other words, false negatives decrease in all cases with respect to the baseline model, implying that far fewer students are now predicted to finish but actually end up dropping out.

As can be seen, an increase in recall is always associated with a decrease in precision and vice versa. In this sense, precision decreases significantly in relation to the base model for all techniques, indicating that more students are predicted to drop out of school but then actually finish (false positives).

On the other hand, the fact that precision decreases, minimizes the chances of overfitting.

It can be seen that, in general, the best classifier is the neural network, especially with SMOTE_SVM resampling, with which a good recall percentage is achieved and with a threshold very similar to the default of 0.5 (Figure 7 and Table 16). This is corroborated by the f1-Score values and the percentage of 84% under the AUC curve, much higher than the rest of the models.

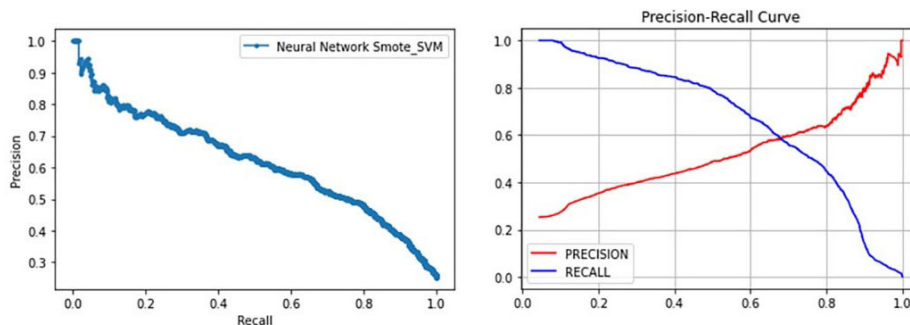


Fig. 7. Precision-Recall (left) and Precision-Recall vs. Threshold (right) curves for the neural network model with the SMOTE_SVM resampling technique

Likewise, from the point of view of preserving the balance between false negatives and false positives, Random Forest with hyperparameters obtains a result very similar to that obtained by neural networks with data distribution alteration techniques, which confirms its suitability for the treatment of imbalanced data.

It is interesting to note the significant improvement of the Logistic Regression model with the incorporation of its own balancing algorithm, achieving a significant increase in recall in relation to the base model, practically with an optimum prediction probability threshold similar to that established by default.

In relation to the f1-Score metric, for the different classifiers, it can be observed that the values are in a very similar range in relation to the optimized models.

Finally, as can be seen, the addition of hyperparameters to the Random Forest classifier, either combined with resampling or without altering the data distribution, significantly improved the model.

3.5 Finding the consensus optimal threshold of predictive probability of college dropout

A representation of the set of components of the f1-Score. vs. Threshold vector for the mean prediction probabilities and subsequent ranking of the three selected models is illustrated in Figure 8.

As demonstrated, the maximum f1-Score provided a consensus optimal threshold among the three selected models of 0.463.

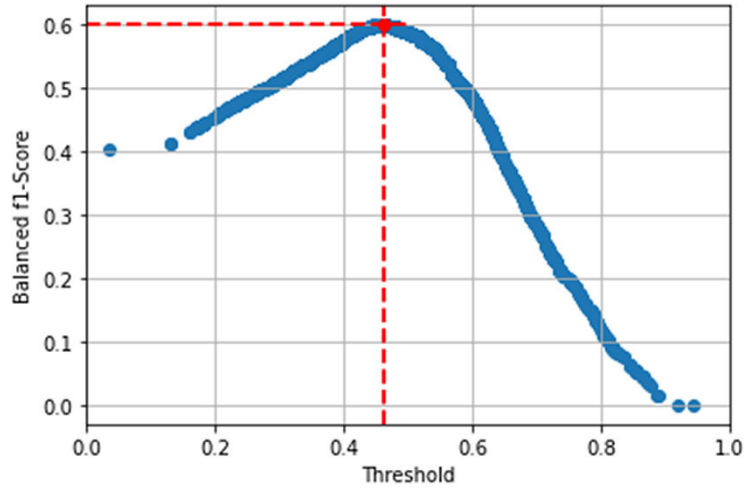


Fig. 8. Determination of the optimal consensus threshold, based on the maximum value of the f1-Score for the three selected models

Table 16 shows a comparison between the different optimal probability thresholds for the base models, including the consensus value.

Table 16. Contrasting optimal prediction probability thresholds for the base models

| Classifier | Threshold | Overall Accuracy | Precision | Recall | f1-Score | ROC/AUC |
|------------------------------|-----------|------------------|-----------|--------|----------|---------|
| Logistic Regression baseline | .50 | .77 | .64 | .24 | .35 | .59 |
| | .243 | .74 | .48 | .65 | .55 | .70 |
| | .463 | .78 | .63 | .26 | .37 | .60 |
| Random Forest baseline | .50 | .79 | .60 | .47 | .53 | .68 |
| | .427 | .78 | .56 | .55 | .55 | .70 |
| | .463 | .78 | .57 | .51 | .54 | .69 |
| Neural Network baseline | .50 | .80 | .68 | .42 | .52 | .67 |
| | .332 | .80 | .59 | .61 | .60 | .81 |
| | .463 | .78 | .60 | .42 | .49 | .66 |

As can be seen, the value of the consensus optimal threshold differs significantly from the optimal values found in the cases of the Logistic Regression and Neural Network base models. However, this value is very close to the optimal threshold of the base model of the Random Forest classifier. This shows that the optimal threshold fitting technique in this case is a very good approximation to reality for imbalanced data and, therefore, its direct application avoids the use of resampling techniques, with the consequent complexity, alteration of the data distribution and possible overfitting problems.

An example of this occurs in the neural network model with oversampling SMOTE_SVM (Figure 9), where the image on the left shows an excellent fit with almost no overfitting in the base model, while in the image on the right, with the application of resampling, there is a perceptible, although not very significant, overfitting.

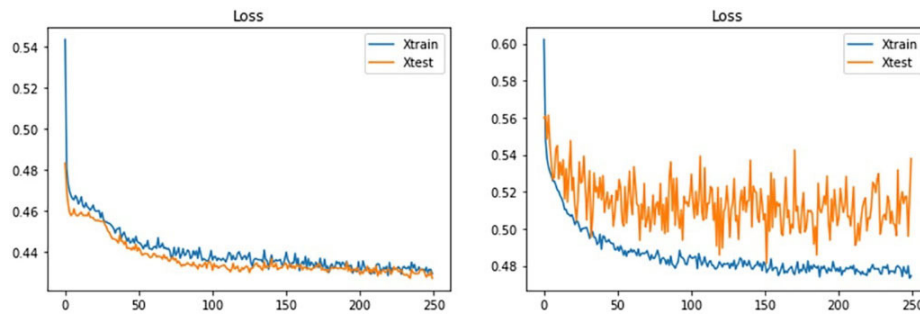


Fig. 9. Creation of overfitting when resampling a base model, in this case with SMOTE_SVM

4 Discussion

In this research article, the need to highlight the importance of considering predictive probability thresholds, in contrast to or in conjunction with the use of other more complex techniques, in a typical scenario of imbalanced data such as the case of late college dropout, was raised. In this sense, although some references have been found on the advantages of implementing this technique, it has not been treated with the required depth and rigor.

In relation to the variables, in this research only those related to the academic context were significant, especially those that had a temporal link with the development of learning, such as the duration of the subjects and the MFP, extensions granted, etc. Although no precedents have been found by authors referring to late dropout, it is coherent that this is the case in this context since other types of variables such as demographic, personal variables or grades [101] have a greater influence on early dropout. These results are supported by most of the educational data mining literature, for example, [102] identified aspects such as student background and age of entry for early dropout; however, [48] concluded that academic success was inferred from both first-year grade point average and the time it took to conclude the degree.

In reference to the imbalanced models of Logistic Regression, Random Forest, and Neural Network, an accuracy of 0.77, 0.79, and 0.80, respectively, was obtained for the default probability threshold of 0.5. This data, which apparently represents a good generalization of the three models, is not so if we look at the rest of the metrics. For example, in the case of Logistic Regression, a low recall value of 0.24 is obtained for the minority class and a high precision of 0.79 for the majority class, which indicates that we are dealing with a typical scenario of imbalanced data. In this context, [70], [100], [103] advise not to rely only on the accuracy value since a significant value can be obtained from this metric, and yet the model fails to recognize the minority class. Consequently, it is important to consider all metrics and establish relationships

between them and the model variables to ensure that a generalizable predictive model is obtained.

For the default threshold and imbalanced data, the best classifier was Random Forest with a recall of 0.47 and an f1-Score of 0.53. The worst performing model was Logistic Regression, with very low metrics of 0.24 and 0.35, for recall and f1-Score, respectively. These values confirm the high robustness of the Random Forest algorithm for imbalanced data as high performance is obtained compared to the rest, without having implemented additional improvement techniques, attributes, or algorithms [70], [104]. These results are supported by authors such as [43], [45] and [105], who consider the Random Forest classifier as one of the most used classifiers ahead of neural networks and logistic regression, besides being intuitive, powerful, and allowing reliable measurements of the variables; on the other hand, Vera et al. [48] found that it is superior to other classification techniques such as Support Vector Machine (SVM), Naive Bayes, or Decision Trees. In short, the Random Forest algorithm clearly wins over other types of classifiers in an imbalanced data scenario and with the default threshold of 0.5, which gives an idea of the ability of this technique to correctly classify all instances of the dataset.

In relation to the adjustment with the optimal probability thresholds of the models without resampling, values well below 0.5 are obtained and, therefore, very significant increases in recall in the cases of Logistic Regression and Neural Network, causing a decrease in precision that, in the case of the Neural Network, manages to reach practically a balance between both metrics. As for the Random Forest classifier, although the decrease in the threshold is much less significant, a good increase in recall is achieved, also achieving a balance with the precision. The robustness of the Random Forest model is therefore demonstrated once again by achieving a balance between false positive and false negative rates with a threshold of 0.427, close to the default value of 0.5. This result is corroborated by authors such as [44], [45], who consider it important to increase recall versus precision to decrease the high cost of the presence of false negatives. In this sense, an increase in recall means that institutions acquire greater efficiency in the probability of predicting abandonment, with repercussions on economic costs, public image, promotions, and government subsidies, among others.

Regarding the fit with the optimal probability thresholds of the models with resampling, in general, the different combinations managed to improve or, at least, maintain (except in some cases) the recall values of the models without resampling. The difference is that the thresholds obtained were between 0.454 and 0.669 ($\bar{x} = 0.56, sd = 0.0649$) for all classifiers, not too far from 0.5 on average. In this context, very significant recall increases were obtained for the Neural Network classifier, reaching a value of 0.75, with an optimal threshold of 0.512 and SMOTE_SVM as resampling technique. This metric, obtained with an optimal threshold very close to 0.5, is considered acceptable when reviewing the f1-Score value of 0.6 and the corresponding area under the curve (AUC) of 0.84, exceeding the value of 0.7 accepted by the research community [70]. These results are, in general, similar to those obtained by [71] with SMOTE-balanced data, with an AUC equal to 0.83, recall of 0.65 and f1-Score of 0.69 for the neural network, and AUC of 0.74, recall of 0.64, and f1-Score of 0.68 for the decision tree.

In order to establish a consensual optimal threshold of probability in the prediction of postgraduate university dropout in the UNEAT and UNINI-MX institutions,

the three best balanced models provided a value of $0.463 \in [0,1]$, less than 0.5, which allowed us to guarantee an increase in the recall metric in relation to the default threshold. This means that it is a reliable threshold because of the high cost of false negatives of the model versus false positives [86], [87], [93].

Contrasting the threshold of the consensus model (0.463) with the thresholds of the base models (without resampling), we matched this threshold to that of the Random Forest classifier (0.427), obtaining very similar metrics. This indicates that a base model Random Forest without resampling, fitted with an optimal threshold, provides a good generalization of the model without the need to introduce noise or to resort to data distribution alteration techniques, which could cause overfitting problems.

5 Conclusions

The work carried out in this research paper enabled a series of relevant conclusions to be drawn, in line with the general objective, on predicting late dropouts from postgraduate distance learning university studies in two educational institutions in the Ibero-American region.

A review of the literature on college dropouts confirmed a majority focus on undergraduate or graduate studies in the face-to-face modality. Only in recent years has there been a growing trend of scientific production on this subject in distance education, albeit always referring to the initial stages of the program where a higher number of dropouts is seen.

On the other hand, the large-scale data generation associated with the academic and administrative processes of educational institutions has led to the obsolescence of analog models and the search for new machine learning tools to complement traditional statistics.

Among these tools, the optimal probability threshold adjustment technique in an unbalanced data scenario, despite its effectiveness, has historically been undervalued against other data resampling methods for unbalanced classes.

The purpose of this paper within this framework, which initiated the research question, was to reduce the gap between the analogical models of face-to-face education and the eLearning context when using machine learning tools for decision making by applying the optimal probability adjustment technique to predict late dropout from graduate university studies, either in isolation or in combination with other techniques, attributes, and algorithms.

The methodology employed provided an affirmative answer by comparing similar metrics between a complex consensus model (threshold of 0.463) and the Random Forests model. Indeed, adjusting the optimal probability thresholds of the base classifiers demonstrated the robustness of this model by achieving a balance between accuracy (0.56) and recall (0.55) with a threshold of 0.427, close to the default value of 0.5. This meant that a base Random Forests model, fitted with an optimal threshold, provided sound results in generalization without the need to resort to data distribution alteration techniques, which could introduce noise into the model and cause overfitting problems. In this sense, the Random Forests classifier proved to be the most robust for unbalanced data and a default threshold of 0.5, with a recall of 0.47 and an f1-Score of 0.53.

Regarding the first research sub-question, the significant variables referred exclusively to the academic setting and not to other social and demographic aspects of the student, whose influence was greater in early dropout. Variables that had an explicit temporal component, such as the duration of the subjects, the duration of the MFP and the extensions, were those that seemed to have the most importance or weight when differentiating between classes.

If the institutional purpose is to deploy preventive strategies aimed at reaching a greater number of students at potential risk of dropping out, sensitivity or recall should be prioritized as an indicator. This ties in with the second research sub-question, as the increase in the recall metric alone predicts far fewer students who will finish, but who actually end up dropping out (false negatives).

In this regard, the top three balanced models that prioritized the recall were: Neural Network with SMOTE_SVM (0.75), Logistic Regression with SMOTE_SVM (0.67) and, being overall better, Random Forests with Hyperparameters (0.6), providing a threshold value of 0.463 by consensus.

In relation to the third research sub-question, the accuracy metric by itself is not a reliable indicator in unbalanced data scenarios. The results obtained of 0.77, 0.79 and 0.80 for Logistic Regression, Random Forests, and Neural Network respectively, and for a default probability threshold of 0.5, do not represent a good approximation of the three models when observing other metrics such as the low value of recall or the f1-Score.

The different tests carried out to answer the fourth research sub-question showed that when adjusting the optimal probability thresholds of the resampled models, thresholds between 0.454 and 0.669 were obtained ($\bar{x} = 0.56, sd = 0.0649$), with a mean not very far from 0.5, but enough to reach a good recall value of 0.75 (threshold of 0.512) for the Neural Network classifier with SMOTE_SVM as the resampling technique.

Lastly, in relation to the values of the overall accuracy metric in the resampled and adjusted models, a narrow range of variation-between 0.71 and 0.78- was observed, meaning that the different machine learning techniques that were used achieved similar results in categorical prediction tasks (classification) when processing data corresponding to the studied variables to predict dropout and study completion.

In summary, the results of this research paper, in accordance with the objective and established research questions, provide a new tool for predicting the late-stage dropout of students pursuing graduate studies online, thus providing an original contribution to the landscape of university studies.

6 Recommendations

Some of the recommendations that could be made in this research work are: increase the population for sampling; consider other types of variables, such as the grades obtained in the subject evaluation phase or the number and quality of the messages exchanged with the MFP advisor; expand the minority class through data collection; search for the best hyperparameters that optimize the different models and, finally, incorporate Application Programming Interfaces (API's) such as Scikeras or pipelines, for greater agility and structuring of the algorithms in Python 3.10.

7 Declarations

Funding: The publication is part of the project “WITH_YOU: Technologies for the dynamic modelling of learners and digital assistants for performance enhancement in e-learning platforms” CPP2021-008349, funded by MCIN/AEI/10.13039/501100011033 and by the European Union-NextGenerationEU/PRTR.

Conflicts of Interest: The authors declare no conflict of interest.

8 References

- [1] J. E. Raffaghelli, and N. Cabrera, “Calidad del eLearning e innovación tecnológica: un proceso en continuo desarrollo,” *Universitat Oberta de Catalunya (UOC)*, 2020. [Online]. Available: <http://hdl.handle.net/10609/125688> [Accessed: July 22, 2022].
- [2] J. Cabero, C. Llorente, and A. Puentes, “La satisfacción de los estudiantes en red en la formación semipresencial,” *Comunicar: Revista científica iberoamericana de comunicación y educación*, vol. 18, no. 35, pp. 149–157, 2010. <https://doi.org/10.3916/C35-2010-03-08>
- [3] C. F. De Oliveira, “How does learning analytics contribute to prevent students’ dropout in higher education: a systematic literature review,” *Big Data and Cognitive Computing*, vol. 5, no. 4, p. 64, 2021. <https://doi.org/10.3390/bdcc5040064>
- [4] E. C. Queiroga, “Uma abordagem para predição de estudantes em risco utilizando algoritmos genéticos e mineração de dados: um estudo de caso com dados de um curso técnico a distância,” In Workshop do VIII Congresso Brasileiro de Informática na Educação (WCBIE), 2019, pp. 119–128 [Online]. <https://doi.org/10.5753/cbie.wcbie.2019.119>
- [5] B. Prenkaj, “A survey of machine learning approaches for student dropout prediction in online courses,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2020. <https://doi.org/10.1145/3388792>
- [6] B. Z. Albreiki, “A systematic literature review of student’ performance prediction using machine learning techniques,” *Education Sciences*, vol. 11, no. 9, p. 552, 2021. <https://doi.org/10.3390/educsci11090552>
- [7] W. Feng, J. Tang, and T. X. Liu, “Understanding dropouts in MOOCs,” In Proceedings of the 23rd American Association for Artificial Intelligence National Conference (AAAI), Honolulu, HI, USA, 27 January–1 February 2019, no. 33, pp. 517–524. <https://doi.org/10.1609/aaai.v33i01.3301517>
- [8] C. Sorensen, “An examination of factors that impact the retention of online students at a for-profit university,” *Online Learning*, vol. 21, no. 3, pp. 206–221, 2017. <https://doi.org/10.24059/olj.v21i3.935>
- [9] A. Hellas *et al.*, “Predicting academic performance: A systematic literature review,” In Proceedings of the Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education, Larnaca, Cyprus, 2–4 July 2018, pp. 175–199. <https://doi.org/10.1145/3293881.3295783>
- [10] L. González, and O. Espinoza, “Calidad en la educación superior: concepto y modelos,” *Calidad en la Educación*, no. 28, pp. 248–276, 2008. <https://doi.org/10.31619/caledu.n28.210>
- [11] P. A. Willging, and S. D. Johnson, “Factors that influence students’ decision to dropout of online courses,” *Journal of Asynchronous Learning Networks*, vol. 13, no. 3, pp. 115–127, 2009. <https://doi.org/10.24059/olj.v13i3.1659>
- [12] I. Lykourantzou, I. Giannoukos, G. Mpardis, V. Nikolopoulos, and V. Loumos, “Early and dynamic student achievement prediction in e-learning courses using neural networks,” *Journal of the American Society for Information Science and Technology*, vol. 60, no. 2, pp. 372–380, 2009. <https://doi.org/10.1002/asi.20970>

- [13] A. P. Rovai, "A practical framework for evaluating online distance education programs," *The Internet and Higher Education*, vol. 6, pp. 109–124, 2003. [https://doi.org/10.1016/S1096-7516\(03\)00019-8](https://doi.org/10.1016/S1096-7516(03)00019-8)
- [14] C. Díaz Peralta, "Modelo conceptual para la deserción estudiantil universitaria chilena," *Estudios pedagógicos (Valdivia)*, vol. 34, no. 2, pp. 65–86, 2008. <https://doi.org/10.4067/S0718-07052008000200004>
- [15] W. L. Cambuzzi, S. J. Rigo, and J. L. Barbosa, "Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach," *Journal of Universal Computer Science*, vol. 21, no. 1, pp. 23–47, 2015.
- [16] V. Tinto, "Limits of theory and practice in student attrition," *Journal of Higher Education*, vol. 53, no. 6, pp. 687–700, 1982. <https://doi.org/10.2307/1981525>
- [17] E. Castaño, S. Gallón, K. Gómez, and J. Vásquez, "Análisis de los factores asociados a la deserción estudiantil en la Educación Superior: un estudio de caso," *Revista de Educación*, vol. 345, pp. 255–280, 2008.
- [18] M. Tan, and P. Shao, "Prediction of student dropout in e-learning program through the use of machine learning method," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 10, no. 1, pp. 11–17, 2015. <https://doi.org/10.3991/ijet.v10i1.4189>
- [19] A. Caison, "Analysis of institutionally specific retention research: a comparison between survey and institutional database methods," *Research in Higher Education*, vol. 48, no. 4, pp. 435–451, 2007. <https://doi.org/10.1007/s11162-006-9032-5>
- [20] A. F. Cabrera, A. Nora, and M. A. Castañeda, "College persistence: structural equations modeling test of an integrated model of student retention," *Journal of Higher Education*, vol. 64, no. 2, pp. 123–139, 1993. <https://doi.org/10.1080/00221546.1993.11778419>
- [21] L. Rodríguez-Muñiz, A. Bernardo, M. Esteban, and I. Díaz, "Dropout and transfer paths: what are the risky profiles when analyzing university persistence with machine learning techniques?" *PLoS One*, vol. 14, no. 6, 2019. <https://doi.org/10.1371/journal.pone.0218796>
- [22] J. S. McCarthy, and S. Earp, "Who makes mistakes? using data mining techniques to analyze reporting errors in total acres operated," *NASS Research Reports 234367*, United States Department of Agriculture, National Agricultural Statistics Service, 2009. <https://doi.org/10.1007/s11162-006-9032-5>
- [23] V. Tinto, "Dropout from higher education: a theoretical synthesis of recent research," *Review of Educational Research*, vol. 43, no. 1, pp. 89–125, 1975. <https://doi.org/10.3102/00346543045001089>
- [24] A. Cano, and J. D. Leonard, "Interpretable multiview early warning system adapted to underrepresented student populations," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 198–211, 2019. <https://doi.org/10.1109/TLT.2019.2911079>
- [25] D. Gašević, S. Dawson, T. Rogers, and D. Gasevic, "Learning analytics should not promote one size fits all: the effects of instructional conditions in predicting academic success," *Internet High. Education*, no. 28, pp. 68–84, 2016. <https://doi.org/10.1016/j.iheduc.2015.10.002>
- [26] A. Bozkurt, M. Yazıcı, and I. E. Aydin, "Cultural diversity and its implications in online networked learning spaces," in *Research Anthology on Developing Effective Online Learning Courses*, Hershey, PA, USA: IGI Global, 2018, pp. 56–81.
- [27] L. Wasserman, "Statistics and Machine Learning," 2012. [Online]. Available: <https://normal-deviate.wordpress.com/2012/06/12/statistics-versus-machine-learning-5-2> [Accessed: July 13, 2022].
- [28] A. B. Bernardo *et al.*, "Predicción del abandono universitario: variables explicativas y medidas de prevención," *Fuentes*, vol. 16, pp. 63–84, 2015. <https://doi.org/10.12795/revistafuentes.2015.i16.03>

- [29] M. Esteban, A. B. Bernardo, and L. J. Rodríguez-Muñiz, "Permanencia en la universidad: la importancia de un buen comienzo," *Aula Abierta*, vol. 44, pp. 1–6, 2016. <https://doi.org/10.1016/j.aula.2015.04.001>
- [30] E. Ghignoni, "Family background and university dropouts during the crisis: the case of Italy," *Higher Education*, vol. 73, no. 1, pp. 127–151, 2017. <https://doi.org/10.1007/s10734-016-0004-1>
- [31] M. V. Santelices, X. Catalán, D. Kruger and C. Horn, "Determinants of persistence and the role of financial aid: lessons from Chile," *Higher Education*, vol. 71, no. 3, pp. 323–342, 2016. <https://doi.org/10.1007/s10734-015-9906-6>
- [32] M. Cukusić, Ž. Garača, and M. Jadrić, "Online self-assessment and students' success in higher education institutions," *Comput Education*, vol. 72, pp. 100–109, 2014. <https://doi.org/10.1016/j.compedu.2013.10.018>
- [33] T. R. Andrade, "Active methodology, educational data mining and learning analytics: a systematic mapping study," *Informatics Education*, vol. 20, pp. 171–204, 2021. <https://doi.org/10.15388/infedu.2021.09>
- [34] B. Prenkaj, P. Velardi, G. Stilo, D. Distanti, and S. Faralli, "A survey of machine learning approaches for student dropout prediction in online courses," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1–34, 2021. <https://doi.org/10.1145/3388792>
- [35] T. E. Miller, T. M. Tyree, K. K. Riegler, and C. H. Herreid, "Using a model that predicts individual student attrition to intervene with those who are most at risk," *Educational and Psychological Studies Faculty Publications*, vol. 85, no. 3, pp. 12–19, 2009.
- [36] M. R. Beikzadeh, S. Phon-Amnuaisuk, and N. Delavari, "Data mining application in higher learning institutions," *International Journal of Informatics in Education*, vol. 7, no. 1, pp. 31–54, 2008. <https://doi.org/10.15388/infedu.2008.03>
- [37] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: moodle case study and tutorial," *Computers and Education*, vol. 51, no. 1, pp. 368–384, 2008. <https://doi.org/10.1016/j.compedu.2007.05.016>
- [38] E. Yukselturk, S. Ozekes, and Y. Türel, "Predicting dropout student: an application of data mining methods in an online education program," *European Journal of Open, Distance and E-Learning*, vol. 17, pp. 118–133, 2014. <https://doi.org/10.2478/eurodl-2014-0008>
- [39] R. Baker, and G. Siemens, "Educational data mining and learning analytics," in *Cambridge handbook of the learning sciences*. Cambridge University Press [online document], 2014, pp. 253–272. <https://doi.org/10.1017/CBO9781139519526.016>
- [40] C. Zhao, and J. Luan, "Data mining: going beyond traditional statistics," *New Directions for Institutional Research*, vol. 131, no. 2, pp. 7–16, 2006. <https://doi.org/10.1002/ir.184>
- [41] F. Sancho, "Introducción al aprendizaje automático," *Universidad de Sevilla*, 2015. [Online]. Available: <http://www.cs.us.es/~fsancho/?e=75> [Accessed: July 27, 2022].
- [42] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: a review and synthesis," *Telematics and Informatics*, vol. 37, pp. 13–49, 2019. <https://doi.org/10.1016/j.tele.2019.01.007>
- [43] F. Agrusti, G. Bonavolontà, and M. Mezzini, "University dropout prediction through educational data mining techniques: a systematic review," *Journal of E-Learning and Knowledge Society*, vol. 15, no. 3, pp. 161–182, 2019. <https://doi.org/10.20368/1971-8829/1135017>
- [44] A. Behr, M. Giese, H. D. Teguim, and K. Theune, "Early prediction of university dropouts—A Random Forest approach," *Journal of Economics and Statistics (Jahrbuecher fuer Nationaloekonomie und Statistik)*, vol. 240, no. 6, pp. 743–789, 2020. <https://doi.org/10.1515/jbnst-2019-0006>
- [45] C. Beulac, and J. S. Rosenthal, "Predicting university students' academic success and major using Random Forest," *Research in Higher Education*, vol. 60, no. 7, pp. 1048–1064, 2019. <https://doi.org/10.1007/s11162-019-09546-y>

- [46] V. Flores, S. Heras, and V. Julian, "Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education," *Electronics*, vol. 11, no. 3, p. 457, 2022. <https://doi.org/10.3390/electronics11030457>
- [47] J. P. Zaldumbide, and V.C. Párraga, "Systematic mapping study of literature on educational data mining to determine factors that affect school performance," presented at International Conference on Information Systems and Computer Science, Quito, Ecuador, 2018. <https://doi.org/10.1109/INCISCOS.2018.00042>
- [48] M. Vera, A. Freitas, P. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: a predictive modeling approach," *Decision Support Systems*, vol. 115, pp. 36–51, 2018. <https://doi.org/10.1016/j.dss.2018.09.001>
- [49] A. Mayra, and D. Mauricio, "Factors to predict dropout at the universities: a case of study in Ecuador," In Proceedings of IEE Global Engineering Education Conference '18, 2018, pp. 1238–1242. <https://doi.org/10.1109/EDUCON.2018.8363371>
- [50] D. Vila, *et al.*, "Detection of desertion patterns in university students using data mining techniques: a case study," *Communications in Computer and Information Science*, vol. 895, pp. 420–429, 2018. https://doi.org/10.1007/978-3-030-05532-5_31
- [51] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, "Perspectives to predict dropout in university students with machine learning," presented at the 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI 2018), San Carlos, Alajuela Province, Costa Rica, 2018. <https://doi.org/10.1109/IWOBI.2018.8464191>
- [52] L. Zea, Y. Piñeros, and J. Rodríguez, "Machine learning for the identification of students at risk of academic desertion," *Communications in Computer and Information Science*, vol. 1011, pp. 462–473, 2019. https://doi.org/10.1007/978-3-030-20798-4_40
- [53] O. Adejo, and T. Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach," *Journal of Applied Research in Higher Education*, vol. 10, no. 1, pp. 61–75, 2018. <https://doi.org/10.1108/JARHE-09-2017-0113>
- [54] M. Nagy, and R. Molontay, "Predicting dropout in higher education based on secondary school performance," In Proceedings IEEE 22nd International Conference on Intelligent Engineering Systems (INES) '18, 2018, pp. 000389–000394, <https://doi.org/10.1109/INES.2018.8523888>
- [55] C. Mason, J. Twomey, D. Wright, and L. Whitman, "Predicting engineering student attrition risk using a probabilistic neural network and comparing results with a backpropagation neural network and logistic regression," *Research in Higher Education*, vol. 59, no. 3, pp. 382–400, 2018. <https://doi.org/10.1007/s11162-017-9473-z>
- [56] M. Alban, and D. Mauricio, "Neural networks to predict dropout at the universities," *International Journal of Machine Learning and Computing*, vol. 9, no. 2, pp. 149–153, 2019. <https://doi.org/10.18178/ijmlc.2019.9.2.779>
- [57] B. Perez, C. Castellanos, and D. Correal, "Applying data mining techniques to predict student dropout: a case study," presented at the IEEE 1st Colombian Conference on Applications in Computational Intelligence (ColCACI) '18, 2018, pp. 1–6. <https://doi.org/10.1109/ColCACI.2018.8484847>
- [58] A. Serra, P. Perchinunno, and M. Bilancia, "Predicting student dropouts in higher education using supervised classification algorithms," In: Gervasi, O., et al. (eds.) ICCSA 2018. LNCS, vol. 10962, pp. 18–33. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-95168-3_2
- [59] K. Y. Diaz, B. Y. Chindoy, and A. A. Rosado, "Review of techniques, tools, algorithms and attributes for data mining used in student desertion," *Journal of Physics: Conference Series*, vol. 1409, no. 1, p. 012003, 2019. <https://doi.org/10.1088/1742-6596/1409/1/012003>
- [60] D. Delen, "Predicting Student Attrition with Data Mining Methods," *Journal of College Student Retention: Research, Theory and Practice*, vol. 13, no. 1, pp. 17–35, 2011. <https://doi.org/10.2190/CS.13.1.b>

- [61] A. Hadad, D. Evin, and B. Drozdowicz, "Modelo para el tratamiento de datos no balanceados basado en redes neuronales autoorganizadas," XVII Congreso Argentino de Bioingeniería y VI Jornadas de Ingeniería Clínica SABI2009, Rosario, Santa Fe, Argentina, 2009.
- [62] A. D. Pozzolo, "Adaptive machine learning for credit card fraud detection," Ph.D. thesis, Université Libre de Bruxelles, 2015.
- [63] F. Fióla, C. E. Alvez, and C. I. Chesñevar, "Un primer acercamiento a un modelo predictivo ajustable por umbrales para detección de fraudes financieros," in *2020 XXI Simposio Argentino de Inteligencia Artificial (ASAI 2020)—JAIIO 49 (Modalidad virtual)*, pp. 114–127 [Online]. Available: <http://sedici.unlp.edu.ar/handle/10915/116433> [Accessed: July 29, 2022].
- [64] Z. Chen, W. Yu, and L. Zhou, "ADASYN—Random forest based intrusion detection model," in *2021 4th International Conference on Signal Processing and Machine Learning*, pp. 152–159 [Online]. <https://doi.org/10.1145/3483207.3483232>
- [65] Chun-Yang P., and You-Jin P., "A new hybrid under-sampling approach to imbalanced classification problems," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1–18, 2021. <https://doi.org/10.1080/08839514.2021.1975393>
- [66] K. Fujiwara K, *et al.*, "Over- and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis," *Front Public Health*, vol. 8, p. 178, 2020. <https://doi.org/10.3389/fpubh.2020.00178>
- [67] I. Tomek, "Two modifications of CNN," In *Proceedings IEEE Transactions on Systems Man and Communications* '76, 1976, pp. 769–772. <https://doi.org/10.1109/TSMC.1976.4309452>
- [68] T. Elhassan, and M. Aljurf, "Classification of imbalance data using Tomek Link (T-Link) combined with Random Under-Sampling (RUS) as a data reduction method," *Global Journal of Technology & Optimization*, p. 111, 2017. <https://doi.org/10.21767/2472-1956.100011>
- [69] M. Kamaladevi, V. Venkataraman, and K. R. Sekar, "Tomek link undersampling with stacked ensemble classifier for imbalanced data classification," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4. [Online]. Available: <https://www.annalsofrscb.ro/index.php/journal/article/view/2751/2283> [Accessed: July 22, 2022].
- [70] S. Dass, K. Gary, and J. Cunningham, "Predicting Student Dropout in Self-Paced MOOC Course Using Random Forest Model," *Information*, vol. 12, no. 11, p. 476, 2021. <https://doi.org/10.3390/info12110476>
- [71] M. A. Miranda, and J. Guzmán, "Analysis of dropouts of university students using data mining techniques," *Formación universitaria*, vol. 10, no. 3, pp. 61–68, 2017. <https://doi.org/10.4067/S0718-50062017000300007>
- [72] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018. <https://doi.org/10.1613/jair.1.11192>
- [73] D. Elreedy, and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32–64, 2019. <https://doi.org/10.1016/j.ins.2019.07.070>
- [74] H-Y. Wang, "Combination approach of SMOTE and biased-SVM for imbalanced datasets," In *Proceedings IEEE International Joint Conference on Neural Networks '08 (IEEE World Congress on Computational Intelligence)*, 2008, pp. 228–231. <https://doi.org/10.1109/IJCNN.2008.4633794>
- [75] W. Jia-Bao, Z. Chun-An, and F. Guang-Hui, "AWSMOTE: an SVM-based adaptive weighted smote for class-imbalance learning," *Scientific Programming*, vol. 2021, 18 pp. 2021. <https://doi.org/10.1155/2021/9947621>
- [76] J. Mathew, M. Luo, C. K. Pang, and H. L. Chan, "Kernel-based SMOTE for SVM classification of imbalanced datasets," In *Proceedings of the IECON 2015 41st Annual Conference of the IEEE Industrial Electronics Society, Yokohama, Japan, 2015*, pp. 001127–001132. <https://doi.org/10.1109/IECON.2015.7392251>

- [77] H. He, Y. Bai, E. A. García, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," In Proceedings of the 2008 IEE International Joint Conference on Neural Networks (IJCNN'08), pp. 1322–1328. <https://doi.org/10.1109/IJCNN.2008.4633969>
- [78] J. Brandt, and E. Lanzén, "A comparative review of SMOTE and ADASYN in imbalanced data classification," 2020. [Online]. Available: <https://www.diva-portal.org/smash/get/diva2:1519153/FULLTEXT01.pdf> [Accessed: July 13, 2022].
- [79] M. Koziarsky, "CSMOUTE: Combined synthetic oversampling and undersampling technique for imbalanced data classification," 2021. [Online]. Available: <https://arxiv.org/pdf/2004.03409.pdf> [Accessed: July 30, 2022]. <https://doi.org/10.1109/IJCNN52387.2021.9533415>
- [80] E. AT, M. Aljourf, F. Al-Mohanna, and M. Shoukri, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method," *Global Journal of Technology and Optimization*, no. S1, pp. 1–11, 2017. <https://doi.org/10.4172/2229-8711.S1111>
- [81] D. Passos, and P. Mishra, "A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks," *Chemometrics and Intelligent Laboratory Systems*, vol. 223, 2022. <https://doi.org/10.1016/j.chemolab.2022.104520>
- [82] M. Kiran, and M. Ozyildirim, "Hyperparameter tuning for deep reinforcement learning applications," *arxiv*, vol. abs/2201.11182, 2022. <https://doi.org/10.48550/arXiv.2201.11182>
- [83] E. Elgeldawi, A. Sayed, A. R. Galal, and A. M. Zaki, "Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis," *Informatics*, vol 8, no. 4, p. 79, 2021. <https://doi.org/10.3390/informatics8040079>
- [84] N. Thai-Nghe, T. Do, and L. Schmidt-Thieme, "Learning optimal threshold for Bayesian posterior probabilities to mitigate the class imbalance problem," *Journal of Science and Tecnology*, vol. 48, no. 4, pp. 38–50, 2010.
- [85] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal thresholding of classifiers to maximize f1 measure," *Machine Learning and Knowledge Discovery in Databases*, vol. 8725, pp. 225–239, 2014. https://doi.org/10.1007/978-3-662-44851-9_15
- [86] Q. Zou, X. Sifa, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Research*, vol. 5, pp. 2–8, 2016. <https://doi.org/10.1016/j.bdr.2015.12.001>
- [87] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "GHOST: Adjusting the decision threshold to handle imbalanced data in machine learning," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2623–2640, 2021. <https://doi.org/10.1021/acs.jcim.1c00160>
- [88] M. Martinsuo, and M. Huemann, "Designing case study research," *International Journal of Project Management*, vol. 39, no. 5, pp. 417–421, 2021. <https://doi.org/10.1016/j.ijproman.2021.06.007>
- [89] M. Azcona, F. Manzini, and J. Dorati, "La unidad de análisis y la unidad de observación," *Ficha de cátedra*, 2019.
- [90] T. M. Barros, P. A. Souza Neto, I. Silva, and L. A. Guedes, "Predictive models for imbalanced data: a school dropout perspective," *Education Sciences*, vol. 9, no. 4, p. 275, 2019. <https://doi.org/10.3390/educsci9040275>
- [91] R. Hernández, C. Fernández, and P. Baptista, "Definiciones de los enfoques cuantitativo y cualitativo, sus similitudes y diferencias," in *Metodología de la investigación*, México, México D.F: McGraw-Hill, 2014, pp. 2–20 [Online]. Available: <https://www.uca.ac.cr/wp-content/uploads/2017/10/Investigacion.pdf> [Accessed: Agost 20, 2022].
- [92] F. J. Pérez, P. Martínez, and M. Martínez, "Satisfacción del estudiante universitario con la tutoría. Diseño y validación de un instrumento de medida," *Estudios sobre educación*, vol. 29, pp. 81–101, 2015. <https://doi.org/10.15581/004.29.81-101>

- [93] J. Redondo, "Comparativa de modelos de Bosques Aleatorios y Redes Neuronales aplicados al mantenimiento predictivo con valores ausentes y datos desbalanceados," trabajo fin de master, Universidad Politécnica de Madrid, 2021 [Online]. Available: https://eprints.ucm.es/id/eprint/68457/1/Javier_Redondo_TFM.pdf [Accessed: Agost 20, 2022].
- [94] R. A. Mohammed, K. W. Wong, M. F. Shiratuddin, and X. Wang, "Scalable machine learning techniques for highly imbalanced credit card fraud detection: a comparative study," In Proceedings 15th Pacific Rim International Conference on Artificial Intelligence '18, 2018, pp. 237–246. https://doi.org/10.1007/978-3-319-97310-4_27
- [95] B. Abella, "Mejora de las predicciones en muestras desbalanceadas", trabajo fin de grado, Universidad Autónoma de Madrid, 2021 [Online]. Available: <http://hdl.handle.net/10486/697900> [Accessed: Agost 18, 2022].
- [96] C. G. Weng, and J. Poon, "A New Evaluation Measure for Imbalanced Datasets," In Proceedings of the 7th Australasian Data Mining Conference (AusDM '08), 2008, vol. 87, pp. 27–32.
- [97] C. Chen, A. Liaw, and L. Breiman, "Using Random Forest to Learn Imbalanced Data," University of California, Tech. Rep Se.; no 666, 2004 [Online]. Available: <https://statistics.berkeley.edu/tech-reports/666> [Accessed: Agost 18, 2022].
- [98] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed. Hoboken, New Jersey: Wiley, 2014 [E-book] Available: <http://www.ccas.ru/voron/download/books/machlearn/kuncheva04combining.pdf> [Accessed: Agost 18, 2022]. <https://doi.org/10.1002/9781118914564>
- [99] M. Guijarro, "Combinación de clasificadores para identificación de texturas en imágenes naturales: nuevas estrategias locales y globales," tesis doctoral, Universidad Complutense de Madrid, 2009 [Online]. Available: <https://eprints.ucm.es/id/eprint/10259/1/T31475.pdf> [Accessed: Agost 18, 2022].
- [100] A. Algarni, "Data mining in education," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, pp. 456–461, 2016. <https://doi.org/10.14569/IJACSA.2016.070659>
- [101] I. M. Khan, A. R. Ahmad, N. Jabeur, and M. N. Mahdi, "Machine learning prediction and recommendation framework to support introductory programming course," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 17, pp. 42–59, 2021. <https://doi.org/10.3991/ijet.v16i17.18995>
- [102] K. B. Eckert, and R. Suénaga, "Análisis de deserción-permanencia de estudiantes universitarios utilizando técnica de clasificación en minería de datos," *Formación Universitaria*, vol. 8, no. 5, 2015. <https://doi.org/10.4067/S0718-50062015000500002>
- [103] W. Koehrsen, "When accuracy isn't enough, use precision and recall to evaluate your classification model," July, 2022 [Online]. Available: <https://builtin.com/data-science/precision-and-recall> [Accessed: Agost 21, 2022].
- [104] K. J. Archer, and R. V. Kimes, "Empirical characterization of Random Forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, pp. 2249–2260, 2008. <https://doi.org/10.1016/j.csda.2007.08.015>
- [105] T. Hamim, F. Benabbou, and N. Sael, "Survey of machine learning techniques for student profile modeling," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 16, no. 4, pp. 136–151, 2021. <https://doi.org/10.3991/ijet.v16i04.18643>

9 Authors

Carmen Lili Rodríguez Velasco has been teaching at the European University of the Atlantic since 2018 and collaborating with the International University do Cuanza in Angola since 2022. She holds a Doctorate in Education, specifically in the line of research in technology and educational innovation with ICT. Her professional career has always been related to research centers, universities, and foundations focused on teaching.

Eduardo García Villena has been teaching since 2015 at the Higher Polytechnic School of the European University of the Atlantic. He holds a Doctorate in Project Engineering since 2011 and is the coordinator of the Department of Environment and Sustainability at the International Iberoamerican University.

Julián Brito Ballester has pursued her professional career in academic coordination, development, training, and personnel selection at the European University of the Atlantic. She holds a Doctorate in Education, specifically in the line of research in occupational competence, education, and employment. She is currently director of the International Office of Quality and Studies of the Iberoamerican University Foundation.

Frigdiano Álvaro Durántez Prados is Doctor Europeus and winner of the Extraordinary Doctorate Award in Political Science from the Complutense University of Madrid. He currently serves as Director of Institutional Relations of the Iberoamerican University Foundation, and is a professor at the European University of the Atlantic.

Eduardo Silva Alvarado holds a Doctorate in Projects from the International Iberoamerican University and a Master's Degree in International Business Management from the Autonomous University of Barcelona. His professional experience with universities is related to advising on the signing of institutional agreements in different countries throughout Europe, Latin America, and the United States of America.

Jorge Crespo Álvarez holds a Doctorate in Civil Engineering from the University of Cantabria and is director of postgraduate and Master's degree programs in Strategic Management in Information Technologies from the European University of the Atlantic. He currently teaches numerous university degrees in the fields of mathematics and statistics.

Article submitted 2022-08-21. Resubmitted 2022-12-12. Final acceptance 2022-12-15. Final version published as submitted by the authors.