López-García, Aarón; Blasco-Blasco, Olga; Liern-García, Marina; Parada-Rico, Sandra E.

Article

Early detection of students' failure using Machine Learning techniques

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: López-García, Aarón; Blasco-Blasco, Olga; Liern-García, Marina; Parada-Rico, Sandra E. (2023) : Early detection of students' failure using Machine Learning techniques, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 11, pp. 1-11, https://doi.org/10.1016/j.orp.2023.100292

This Version is available at:
https://hdl.handle.net/10419/325777

Operations Research Perspectives 11 (2023) 100292

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

Early detection of students' failure using Machine Learning techniques

Aarón López-García [a], Olga Blasco-Blasco [b,*], Marina Liern-García [c], Sandra E. Parada-Rico [d]

[a] Universidad Politècnica de Valencia, Camí de Vera, s/n, Valencia, 46022, Spain
[b] Departamento de Economía Aplicada, Facultad de Economía, Universitat de València, Avda. Tarongers s/n, Valencia, 46022, Spain
[c] Universitat de València, Av. Blasco Ibáñez 13, Valencia, 46010, Spain
[d] School of Mathematics, Industrial University of Santander, Carrera 27 Calle 9, Edificio Camilo Torres, Bucaramanga, 680002, Colombia

ARTICLE INFO

Keywords:
Student performance
Academic failure
XGBoost
Gradient Boosting Machine
UW-TOPSIS
ADASYN

ABSTRACT

The educational system determines one of the significant strengths of an advanced society. A country with a lack of culture is less competitive due to the inequality suffered by its people. Institutions and organizations are putting their efforts into tackling that problem. Nevertheless, it is not an easy task to ascertain why their students have failed or what are the conditions that affect such situations. In this work, an intelligent system is proposed to predict academic failure by using student information stored by the Industrial University of Santander (Colombia). The prediction model is powered by the XGBoost algorithm, where a TOPSIS-based feature extraction and ADASYN oversampling have been conducted. Hyperparameters of the classifier were tuned by a cross-validated grid-search algorithm. We have compared our results with other decision-tree classifiers and displayed the feature importance of our intelligent system as an explainability phase. In conclusion, our intelligent system has shown a superior performance of our prediction model and has indicated to us that economic, health and social factors are decisive for the academic performance of the students.

1. Introduction

Superior education, often referred to as higher education, is indispensable for the holistic development of a country. The growth and development of social, economic, and cultural features rely on it. For that reason, advanced societies are doing their utmost to carry out actions to promote the cultural enrichment of their population. Human capital development is essential for the innovation and research of a country. Therefore, a large part of the country's strategies have to be aligned with access to quality higher education.

With regard to public universities, the policies taken by the main responsible play a crucial role in providing accessible and quality education. The implementation of strategic plans has to be realized by taking into account multiple approaches such as curriculum development, adequate funding, sustainability, diversity, or inclusion among others. It is easy to understand that these concepts can significantly improve the situation of public universities. However, the management of these institutions depends entirely on the public investment in research and development of the country. In the particular case of Colombia, in 2020 the country's R&D expenditure was 0.29% of its GDP, while the average for Latin America and the Caribbean was 0.69%. As a comparison, the European Union invested 2.32% and the United States 3.45% of their respective GDPs [1].

One of the biggest problems in achieving the cultural enrichment goal is university failure and dropout rates. All the efforts made by public administrations are useless if no measures are taken to address this fact. This concern is even more acute in Latin America, where the dropout rate is sometimes as high as 50%, with 39% in countries such as Colombia [2]. So, we need tools for intelligent decision support. Not only in order to reach as many students as possible, but also because trying to do this task manually would be impossible. Moreover, such intelligent systems cannot be designed as black boxes since both the heads of the institution and the technical team staff need to be familiar with the functioning of these tools.

In this work, the main objective is to predict academic failure using the data available by the Academic Excellence Support of the Industrial University of Santander (Colombia), whose acronym is SEA-UIS because of the Spanish name "*Sistema de Excelencia Académica de la Universidad Industrial de Santander*". For this purpose, we have analyzed the given dataset, in particular, the situation of the students before the beginning of the first course. In this way, the stakeholders can take appropriate action as soon as students enter college and thus take strategic measures to prevent further failure.

With the aim of being able to predict academic failure in the Colombian university, we have generated a decision support system

* Corresponding author.
E-mail addresses: logara8@alumni.uv.es (A. López-García), olga.blasco@uv.es (O. Blasco-Blasco), lierngar@alumni.uv.es (M. Liern-García), sanevepa@uis.edu.co (S.E. Parada-Rico).

https://doi.org/10.1016/j.orp.2023.100292
Received 26 April 2023; Received in revised form 12 September 2023; Accepted 18 November 2023
Available online 20 November 2023

powered by an XGBoost classifier [3]. The main reason we have selected the XGBoost technique is that its structure is composed of decision trees, which is an easy-to-understand machine-learning technique. In this manner, we can guarantee transparency to the managers of the institution. Another underlying advantage of decision trees is that their explainability is inherent [4]. Then, we can give significant information to the stakeholders about the strengths and weaknesses of the university. Apart from having the information given by the SEA-UIS, we have performed a Multiple-Criteria Decision Making (MCDM) feature extractor guided by TOPSIS-based features [5]. In addition, we have handled the imbalanced problem of our dataset with the ADASYN technique [6]. Finally, we have implemented other techniques to contrast the results obtained and check the consistency of each of the parts of our system.

The remainder of this paper is structured as follows. In Section 2, we include a brief literature review. In Section 3, we present the methodology implemented for this work, composed of the five stages required for performing the classification task. In Section 4, we introduce the case study of the Industrial University of Santander (Colombia) with a description of the problem and the strategy conducted. In Section 5, we show the results obtained in our case study and we discuss the main contributions of the project. Finally, Section 6 concludes this paper with the main conclusions of the work.

2. Literature review

The analysis of student performance is a topic that is well-studied by both institutions and research teams. The study of this subject is a great challenge not only because of the importance of approaching this problem but also because of the difficulty of generating appropriate methodologies for each particular problem. Many factors determine the academic performance of students. Hence, all of them must be taken into account when predicting how their evolution will be during their time at university. In most of the studies conducted, the approach has been conducted from three non-exclusive perspectives.

First, we found the direct use of indicators for summarizing the multidimensional information of data [7,8]. This kind of study has been quite implemented at SEA-UIS (Parada et al. [9,10],Liern et al. [11]). The main counterpart attached to this approach is that it is only applicable to posterior evaluation. Then, the use of indicators is not valid for making predictions about possible student outcomes. Another problem that arises when constructing synthetic indicators is the aggregation technique and the weighting scheme attached to it [12].

Second, the implementation of statistical methods is also a recurrent approach. The analysis of variance (ANOVA) models gives us relevant information about the different groups of students [13,14]. For instance, Mushtaq and Khan [15] showed that linear regression analysis can be used to determine the correlation among factors affecting students' academic performance. Although statistical methods give us significant information about the generalizations and associations among students, we just know the global situation of past students. As a consequence, we cannot make inferences about future events, thus complicating the early detection of school failure.

Third, we found the use of data mining and machine learning techniques. This approach turns out to be more appropriate for studying academic performance since the information that is gathered a priori is utilized to make a posteriori case study. The objective is to create an intelligent system able to learn from the dataset of a given university so that we can recognize and generalize common patterns among students. The combination of supervised machine learning techniques is commonly proposed for model benchmarking [16–19]. In addition, we can also find sophisticated strategies for facing such a task. For example, Hidayah et al. [20] used neuro-fuzzy systems, and [21] implemented an artificial recurrent neural network. It is noteworthy to mention the efforts made to study the gender gaps that may occur within educational centers, thus guaranteeing gender equality. An example with that line is conducted by Sapiezynski et al. [22], where they applied class performance with Linear Discriminant Analysis (LDA) via an imbalanced gender perspective. Although machine learning techniques are very effective for predicting students' performance, it is not an easy task to understand the inner functioning of a great part of the supervised techniques. Then, sometimes it is preferred to reduce the model's performance and use basic methods that ensure explainability and transparency. That is why decision trees (DTs) are usually selected for approaching academic failure prediction. In addition, their effectiveness has been shown over small datasets [23]. For the failure and/or dropout prediction, the tree ensemble methods turned out to be very useful as well as extreme gradient boosting algorithms. Their performance has been proven in such prediction tasks, showing that hyperparameter tuning considerably improves the outcomes [24,25].

Another point to highlight is the emphasis made on marginal groups. When working with real data of university students we should bear in mind that approximately 70% of them do not have a problem achieving the proper performance. For this reason, it is important to consider data augmentation techniques to generate synthetic samples for the minority classes. Oversampling techniques have been shown as proper strategies to enhance the detection of rare/outlier cases [26]. An example of oversampling methods for tackling imbalanced conditions is found in Thai-Nghe et al. [27], where they applied the SMOTE algorithm [28] to oversample the students' set. Despite the fact these techniques can improve the learning process, not all oversampling algorithms generate data in the same way. As a manner to solve this concern, He et al. [6] introduced the ADAptative SYNthetic (ADASYN) [6] algorithm. It uses weighted distributions for oversampling considering their level of difficulty in learning.

Finally, we want to emphasize the development of well-established datasets with detailed information about different academic, economic, and social attributes. It is paramount for analyzing academic performance over actual data [29,30]. That stage is very important since it gives an experimental scenario for researchers with real conditions. Sometimes the information gathered by the institutions does not contain as many characteristics as we would like due to the sensitivity and privacy of the study subjects. An example of this sort of dataset is found in [31], who published a dataset of students in secondary education at two Portuguese schools. Even though these kinds of datasets are very useful for experimental purposes [25], the use of curated data makes our studies not as credible or scalable. For that purpose, in this study, we only utilized samples of actual students of the SEA-UIS institution.

3. Methodology

In this paper, we have combined multiple approaches from different fields with the aim of predicting academic failure at the Industrial University of Santander. To this end, we have implemented a tree-based classifier so that we can know not only those students who are prone to fail but also understand why this is likely to occur. In other words, we want to remove the black boxes commonly attached to most machine-learning algorithms [32], thus offering greater credibility and transparency to the SEA-UIS. The core idea is to measure the underlying feature importance when predicting. Because the success-to-failure ratio for the institution is disproportionate, such natural imbalance means a problem for the learning procedure. That is why we have implemented an oversampling technique focused on marginal samples to increase the accuracy and incorporate a major generalization. We have depicted the procedure conducted in Fig. 1, where the five different stages involved are remarked.

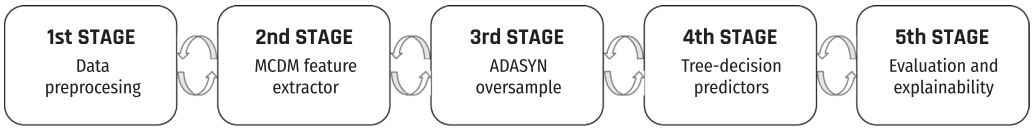

Fig. 1. Methodological steps carried out in this paper to perform the classification of an imbalanced dataset.

3.1. 1st STAGE: Data preprocessing

The dataset used for this work has been transferred by the Industrial University of Santander. Hence, we have followed the guidelines indicated by the UIS-SEA institution in terms of data importation and curation. Regardless of the employed dataset, this stage is crucial in any machine-learning task [33] or knowledge discovery technique [34]. The preprocessing step that deserves to be mentioned is the normalization functions utilized. They were originally proposed by Liern et al. [11], thus measuring the similarity with an ideal fixed by the interval. The mathematical definition of the functions that compute the normalization of the dataset is described in 4.2.

3.2. 2nd STAGE: Multiple Criteria Decision-Making feature extractor

In this section, we briefly describe the Multiple Criteria Decision-Making (MCDM) methods utilized for building additional informative variables. The two methods are the TOPSIS and a generalization of this technique, where the weights do not need to be fixed *a priori*.

3.2.1. TOPSIS method

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [5] is a multiple-criteria decision-making method to rank alternatives regarding positive and negative ideal solutions. Given a weighting scheme, TOPSIS evaluates the dataset by making full use of the selected criteria. For its application, we just need to consider normalization and vector distance. Then it computes the relative proximity to the ideal solutions to provide a cardinal ranking of the alternatives in descending order. It was originally defined with the Euclidean distance and vector normalization as described in the following steps.

1. Determine the decision matrix $[x_{ij}]$, so that $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$, where N and M are the number of alternatives and criteria respectively.
2. Normalize the decision matrix as $[r_{ij}]$, where $r_{ij} \in [0,1]$, $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$.
3. Calculate the weighted normalized decision matrix $[v_{ij}]$ according to the set of weights $\{w_j\}_{j=1}^{M}$ such that $\sum_{j=1}^{M} w_j = 1$. Then, we have $v_{ij} = w_j r_{ij}$, $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$.
4. Positive ideal and negative ideal are determined as $A^+ = (v_1^+,\dots, v_M^+)$ and $A^- = (v_1^-,\dots,v_M^-)$ where

$$v_j^+ = \begin{cases} \max\limits_{1\le i\le N} v_{ij} & \text{if} \quad j \in J_{max} \\ \min\limits_{1\le i\le N} v_{ij} & \text{if} \quad j \in J_{min} \end{cases} \quad 1 \le j \le M,$$
$$v_j^- = \begin{cases} \min\limits_{1\le i\le N} v_{ij} & \text{if} \quad j \in J_{max} \\ \max\limits_{1\le i\le N} v_{ij} & \text{if} \quad j \in J_{min} \end{cases} \quad 1 \le j \le M, \tag{1}$$

   where J_{max} is the set of criteria to be maximized and J_{min} is the set of criteria to be minimized.
5. Calculate the separation measures with regard to A^+ and A^-,

$$D_i^+ = \sqrt{\sum_{j=1}^{M}(v_{ij} - v_j^+)^2} \text{ and } D_i^- = \sqrt{\sum_{j=1}^{M}(v_{ij} - v_j^-)^2}\ ,\ 1 \le i \le N. \tag{2}$$

6. Calculate the relative proximity to the ideal solutions using the relative index:

$$R_i = \frac{D_i^-}{D_i^+ + D_i^-}\ ,\ 1 \le i \le N. \tag{3}$$

7. Rank the alternatives by descending order of the values $\{R_i\}_{i=1}^{N}$.

In such a way, the TOPSIS technique transforms a set of features $N \times M$ dimensional into a decision vector in $[0,1]^N$ that yields decisive information about the global situation of the alternatives involved. Moreover, we could modify the step 4 in which we get $A^+ = (1,\dots,1)$ and $A^- = (0,\dots,0)$. Then, the result is a rank reversal methodology so that it avoids results that are dependent on data [35].

3.2.2. Unweighted TOPSIS

The unweighted TOPSIS (UW-TOPSIS) technique [36] ranks decision alternatives based on the classical TOPSIS approach mentioned in 3.2.1, however, this method does not require the introduction of a priori weights. As a result, the method does not consider the relative importance of the criteria, but it fits the weights in order to solve the optimization problems (minimize/maximize) that involve the relative proximity function (R) in step 6. The output gives us information about both minimal and maximal possible rank values per each alternative. Additionally, we can utilize the optimal weights obtained to evaluate the sensitivity of the method against such intervals. Unlike the weighted approach, this algorithm can be conducted by following the next six steps, in which step 3 is removed.

1. Determine the decision matrix $[x_{ij}]$, so that $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$, where N and M are the number of alternatives and criteria, respectively.
2. Normalize the decision matrix as $[r_{ij}]$, where each $r_{ij} \in [0,1]$ for each $i \in \{1,\dots,N\}$ and $j \in \{1,\dots,M\}$.
3. Positive ideal and negative ideal are determined as $A^+ = (v_1^+,\dots, v_M^+)$ and $A^- = (v_1^-,\dots,v_M^-)$ where

$$v_j^+ = \begin{cases} \max\limits_{1\le i\le N} r_{ij} & \text{if} \quad j \in J_{max} \\ \min\limits_{1\le i\le N} r_{ij} & \text{if} \quad j \in J_{min} \end{cases} \quad 1 \le j \le M,$$
$$v_j^- = \begin{cases} \min\limits_{1\le i\le N} r_{ij} & \text{if} \quad j \in J_{max} \\ \max\limits_{1\le i\le N} r_{ij} & \text{if} \quad j \in J_{min} \end{cases} \quad 1 \le j \le M, \tag{4}$$

   where J_{max} is the set of criteria to be maximized and J_{min} is the set of criteria to be minimized.
4. Given $\Omega = \left\{w = (w_1,\dots,w_M) \in [0,1]^M : \sum_{j=1}^{M} w_j = 1\right\}$ and d a distance defined in $[0,1]^M \times [0,1]^M$, we consider the separating functions $D_i^+, D_i^- : \Omega \to [0,1]$ for each of $i \in \{1,\dots,N\}$ as

$$D_i^+(w) = d\Big((w_1 r_{i1},\dots,w_M r_{iM}),(w_1 v_1^+,\dots,w_M v_M^+)\Big), \tag{5}$$

$$D_i^-(w) = d\Big((w_1 r_{i1},\dots,w_M r_{iM}),(w_1 v_1^-,\dots,w_M v_M^-)\Big). \tag{6}$$

5. The relative proximity function to the ideal solutions is defined as $R_i : \Omega \to [0,1]$ with:

$$R_i(w) = \frac{D_i^-(w)}{D_i^+(w) + D_i^-(w)}, 1 \le i \le N. \tag{7}$$

6. For each $i \in \{1, \dots, N\}$, the values R_i^L and R_i^U are calculated when solving mathematical programming problems over R_i considering the set of weights as the problem variables

$$R_i^L = \min_{1 \le i \le N} \left\{ R_i(w) : \sum_{j=1}^{M} w_j = 1, l_j \le w_j \le u_j \right\}, 1 \le i \le N. \quad (8)$$

$$R_i^U = \max_{1 \le i \le N} \left\{ R_i(w) : \sum_{j=1}^{M} w_j = 1, l_j \le w_j \le u_j \right\}, 1 \le i \le N. \quad (9)$$

Where l_j lower bound and u_j upper bound within w_j, $\forall j \in \{1, \dots, M\}$.

The main consequence of the absence of weights is that the Eqs. (2) and (3) in the classical TOPSIS are converted into functions that depend on the $w \in \Omega$ as in Eqs. (5), (6), and (7). Hence, the output is not only a decision interval $[R_i^L, R_i^U]$ but a set of optimal weights $\{w_i^{*L}, w_i^{*U}\}$ which give us information about the relative importance of the criteria per alternative. It is worth mentioning that uwTOPSIS does not rank alternatives per se. If our goal were to sort a number of alternatives, we would need to define a rank function that involves the values of R_i^L and R_i^U. In order to facilitate the applicability of this MCDM technique, we have designed a GitHub repository[1] [37] written in Python programming language.

3.3. 3rd STAGE: ADASYN oversample

ADAptive SYNthetic (ADASYN) sampling approach [6] is an oversampling algorithm for learning from imbalanced datasets. The underlying idea of this technique is the use of a weighted distribution over the minority class examples according to their difficulty in learning. Then, ADASYN generates synthetic elements of data that reduce the learning bias and the decision boundary of such samples.

Considering the dataset $\mathcal{D} = \{X_i, y_i\}_{i=1}^N$, we distinguish the minority (m) and majority (M) samples as N_m and N_M respectively, so that, $N_m \le N_M$ and $N_m + N_M = N$. Then, the ADASYN oversampling technique may be applied following the steps, in which we can assume, without limiting the generality of the foregoing, that the dataset is sorted so that the first N_m elements belong to the minority class.

1. Define $G = (N_M - N_m)\mu$ as the number of elements to generate, with $0 \le \mu \le 1$ the desired proportion to oversample.
2. Compute the KNN algorithm over the minority class to get the values $\{r_i\}_{i=1}^{N_m}$, where $r_i = \Delta_i / K$ and Δ_i is the cardinality of the neighbors of X_i that belongs to the majority class.
3. Normalize $\{r_i\}_{i=1}^{N_m}$ as $\bar{r}_i = r_i / \sum_{i=1}^{N_m} r_i$ per each $1 \le i \le N_m$.
4. Generate $g_i = \bar{r}_i G$ samples of each $i \in \{1, \dots, N_m\}$ instance, providing a μ-balanced dataset.

As a consequence of the step 3, we notice that $\|\bar{r}\|_1 = 1$, hence we have build a density distribution of $\{r_i\}_{i=1}^{N_m}$. Moreover, a new synthetic dataset can be generated with a partial proportion over the majority class, as stated in step 1. As long as we set $\mu = 1$, we create a fully balanced dataset.

3.4. 4th STAGE: Tree-decision predictors

In the field of Machine Learning, a decision tree (DT) is a tree-like structure that has attached a set of tests that spread over the graph. DT determines decision thresholds that generate the nodes and final leaves by means of splitting algorithms such as the Gini index. They are considered one of the most popular models for both regression and classification, however, they are prone to overfitting and outliers influence [38]. With the aim of preventing such limitations and improving the performance of DTs, Ho [39] designed the Random Forest (RF) model. With the use of stochastic modeling, it was shown that RF is able to increase the accuracy at the same time it is limiting the overfit.

[1] https://github.com/Aaron-AALG/uwTOPSIS

3.4.1. Gradient Boosting Machines

Gradient Boosting Machines (GBM) is a set of algorithms that learns from data via weak learners through an additive approach [40]. The underlying idea relies on the iterative search of an approximation function that accurately maps the response of a given training data samples $\mathcal{D} = \{\mathbf{X}, y\} = \{X_i, y_i\}_{i=1}^N$ of known values. Regarding a loss function $\mathcal{L}(y, \cdot)$, GBMs estimate the objective function of the minimization problem that implies the fitting of the joint distribution of (X_i, y_i)-pairs. In turn, the optimal function may be defined as in Eq. (10):

$$F^*(\mathbf{X}) = \underset{F}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{X},y} \mathcal{L}\left(y, F(\mathbf{X})\right). \quad (10)$$

Now, we build the additive expansion of an approximation function $F(\mathbf{X}; \mathbf{a})$ where $\mathbf{a}$ is the set of characteristic parameters. Thus, we can estimate the objective by following the ensemble technique of Eq. (11), i.e. by a weighted accumulation sum of learners.

$$\left.\begin{aligned} F_m(\mathbf{X}) &= F_{m-1}(\mathbf{X}) + \rho_m h_m(\mathbf{X}; \mathbf{a}). \\ F_0(\mathbf{X}) &= \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}(y_i, \alpha). \end{aligned}\right\} \quad (11)$$

Here, h_m are the weak learners and ρ_m are the multipliers of the linear search. The iterative upload steps of the algorithm are described as the procedure described in Eqs. (12). The underlay goal is to greedily improve the model performance by minimizing the loss function at each time (h_m) via linear search (ρ_m).

$$\begin{aligned} \tilde{y}_{i,m} &= -\left[\frac{\partial \mathcal{L}(y, F_{m-1}(X_i))}{\partial F_{m-1}(X_i)}\right]_{i,m}. \\ \mathbf{a}_m &= \underset{\mathbf{a},\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left[\tilde{y}_i - \beta h(X_i, \mathbf{a})\right]. \\ \rho_m &= \underset{\rho}{\operatorname{argmin}} \sum_{i=1}^{N} \mathcal{L}(y_i, F_{m-1}(X_i) + \rho h(X_i; \mathbf{a}_m)). \\ F_m(\mathbf{X}) &= F_{m-1}(\mathbf{X}) + \rho_m h(\mathbf{X}; \mathbf{a}_m). \end{aligned} \quad (12)$$

We can use this procedure to minimize any differentiable loss $\mathcal{L}$ with forward stage-wise additive modeling. It basically fits the learners $h(\mathbf{X}, \mathbf{a})$ to the responses of the approximated member of the pseudoresponses $\tilde{y}_i$ of the loss function in the steepest-descent strategy. For the particular case in which we consider decision trees as base learners h, we are focusing on the Gradient Tree Boosting (GTB) methodology.

3.4.2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) [3] is a GTB ensemble with a scalable end-to-end system. It is based on a similar approach as the gradient boosting additive procedure described in Section 3.4.1. Nonetheless, since XGBoost makes use of decision trees as weak learners, the loss function to be minimized during training is regularized in order to penalize the tree complexity as shown in Eq. (13).

$$\begin{aligned} &\mathcal{L}_{XGB}(y, F(\mathbf{X})) = \mathcal{L}(y, F(\mathbf{X})) + \sum_k \Omega(h_k), \\ &\text{where } \Omega(h_k) = \gamma T + \frac{\lambda}{2} \|w_{h_k}\|_2^2. \end{aligned} \quad (13)$$

In this manner, given a total number of trees (T), Ω helps to avoid over-fitting by smoothing the learned weights (w). The parameter γ regularizes the loss reduction gain and so determines the complexity of the tree-based learners. The λ parameters control the impact of the weights over the loss. It is easy to note that, when $\gamma = \lambda = 0$ in Eq. (13), we have the same methodology as tree-GBMs.

In the training phase, the empirical loss can be controlled through the formal definition of the iterative process. If we account for the constructive sequence that takes place in GBMs (Eq. (11)), then at iteration k we can measure the loss by means of Eq. (14).

$$\mathcal{L}_{XGB}(y, F_k(\mathbf{X})) = \sum_{i=1}^{N} \mathcal{L}\left(y_i, F_{k-1}(X_i) + h_k(X_i)\right) + \sum_k \Omega(h_k). \quad (14)$$

XGBoost is known to be one of the state-of-the-art techniques in machine learning tasks [41]. Its effectiveness and applicability have

been proven in several challenging areas of classification [42] and regression [43], proving an excellent performance with easy and scalable launching. Moreover, XGBoost is also a good classifier in imbalance tasks such as financial assessment [44], credit scoring [45], or face image manipulation [46].

3.5. 5th STAGE: Evaluation and explainability

In order to give credibility to the research conducted, we have to evaluate the selected models and give an explanation about how they perform the predictions. Then, we can give proper feedback to institutions from two points. Firstly, early detection of academic failure and, secondly, assessment of the current situation of their university students. For the model evaluation, we have performed a classical approach to gauge a binary classification using metrics from the confusion matrix. In regard to the model diagnosis, the field of eXplainable Artificial Intelligence (XAI) has received much attention from the part of the research community [47]. In our case, we manage decision tree predictors in which the node structure describes the class distribution of the predicted classes, giving us fast and consistent local explanations [48]. Then, we can extract the attribute relevance when predicting thanks to the transparency presented by itself [49].

4. Case study: Student classification according to its academic performance

In this section, we explain the procedure we followed during our experiments. Since descriptive analysis may play a key role in helping decision-makers at the Industrial University of Santander (UIS), Colombia, it is essential to know how the data is obtained and stored. The more rigorous and accurate we are, the greater we will be able to support the education system. We need to make a brief reference to the higher education system in Colombia before describing the variables to be considered in this paper. It should be noted that the educational system has improved significantly in recent decades and has faced several challenges, among which the expansion of coverage, the strengthening of technical and technological training, and the creation of accreditation and quality institutions stand out [50]. Nevertheless, according to the National Competitiveness Report 2021–2022 published by "El Observatorio de la Universidad Colombiana, OUC [51]", the proportion of people aged 25–34 with higher education in Colombia is 30%, compared to 46% of the OECD average. The coverage rate, which measures the ability of the education system to meet social demand for education, is 53.9% in 2021, a figure well below the OECD average of around 75%. Although this figure is low, it is improving year by year thanks to free fees at public universities, support programs for university students, and the development of a comprehensive strategy to address young people at risk of dropping out of higher education. In this context, and in accordance with the recommendations of the Ministry of Education, the UIS has set up the Academic Excellence Support System (SEA) at its campus in the city of Bucaramanga, Colombia [52]. This service collects and obtains relevant information from students before they enroll at the university. In this study, we consider the sample of students admitted in the first year of science and engineering degrees to predict academic failure.

4.1. Experimental dataset

For this work, we consider the dataset used by Blasco-Blasco et al. [12], provided by SEA, with 3534 students. This dataset comprises academic, cognitive, economic, health, and social criteria. More detailed information on these features can be found in Parada et al. [53] and Parada et al. [9].

We have a dataset with five features that describe the initial conditions of the UIS students. Then, such combinations of variables are supposed to mark the course of events during the first year at the university. In Parada et al. [10], the authors proved that the usage of this set can lead to interesting studies about the academic achievements of the institution via adequacy indicators. Consequently, we have been trying to make the best use of the given data.

Once the semester has finished, the organization stores the final marks of the students to verify whether they have passed the tests. In this case, we analyzed the final grades of calculus and algebra subjects. Both variables ranged between 0 and 5 so it is assumed that a student with an average lower than 2.5 has not passed the mathematical test, and so in this paper, it is considered a failure. Then, we define the binary variable y, which determines whether a student has passed the course or not as shown in Eq. (15).

$$y_i = \begin{cases} 1 & \text{if} \quad \frac{1}{2}(Algebra + Calculus)_i < 2.5 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for all} \quad 1 \leq i \leq N. \tag{15}$$

where $(Algebra + Calculus)_i$ is the sum of the grades in algebra and calculus subjects for the ith individual.

Once we have calculated the y vector, we noticed that the proportion of our sample is $3:7$ with regard to the failure event. To be more specific, the proportion of students that pass the course is 68.90%. Although it is considered good news for the university, due to their great work, it now leads to an imbalanced learning problem. There are many methods to face such problems in which we can distinguish two procedures: undersampling and oversampling. Given that we want to pay full attention to every single element of the sample, we decided to apply an oversampling technique. Moreover, as we wanted to cluster the students regarding their academic failure, we would like to emphasize the $\{y_i = 0\}$-class.

For the above-mentioned reasons, we decided to apply the ADASYN algorithm. As we detailed in 3.3, this method generates copies of those cases hard to learn during training, so we are highlighting the excluded classes. After applying the oversampling technique, the synthetic dataset is now composed of 4969 elements so that 2533 samples are considered as $\{y_i = 1\}$-class. As a result, the percentage of failure now is 50.98%.

4.2. Normalization

It is well known that data must be properly normalized before the computations are applied (see, for instance, Sola and Sevilla [54] or Trebuňa et al. [55]). As we have described in Sections 3.2.1 and 3.2.2, a normalization technique is needed to implement both methods. Therefore, the normalization functions to be used are the same as those implemented by Liern et al. [11], in which the functions η and ξ are given by the following expressions:

$$\eta_{A,a,b,B;k_1,k_2}(x) = \begin{cases} \dfrac{1 - e^{k_1 \frac{x-A}{a-A}}}{1 - e^{k_1}} & \text{if} \quad A \leq x < a, \\ 1 & \text{if} \quad a < x < b, \\ \dfrac{1 - e^{k_2 \frac{B-x}{B-b}}}{1 - e^{k_2}} & \text{if} \quad b < x \leq B, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

$$\xi_{A,a,b,B}(x) = \begin{cases} \dfrac{x-A}{a-A} & \text{if} \quad A \leq x < a, \\ 1 & \text{if} \quad a < x < b, \\ \dfrac{B-x}{B-b} & \text{if} \quad b < x \leq B, \\ 0 & \text{otherwise.} \end{cases}$$

Due to the definition of the criteria, the (A, a, b, B) array represents the trapezoidal fuzzy shape and the (k_1, k_2) coefficients are the left and right exponents that determine the convexity or concavity of the function. Table 1 shows the transformations of the data and the kind of normalization employed per each feature.

Table 1
Criteria, DDBB values, range transformations, ideal elements, and normalization functions.

Criteria	Original	Transformation	Ideals	Normalization
Academic	{VL, L, LM, M, MH, H, VH}	[1, 7]	[6,7]	$\eta_{1,6,7,7;1,0}(x)$
Cognitive	{VL, L, LM, M, MH, H, VH}	[1, 7]	[6, 7]	$\eta_{1,6,7,7;-1,0}(x)$
Economic	[0, 1]	[0, 1]	[0.8, 1]	$\xi_{0,0.8,1,1}(x)$
Health	[0, 0.65]	[0, 0.65]	0.65	$\xi_{0,0.65,0.65,0.65}(x)$
Social	{0.1, 0.5, 0.7, 1}	[0.1, 1]	[0.7, 1]	$\xi_{0.1,0.7,1,1}(x)$

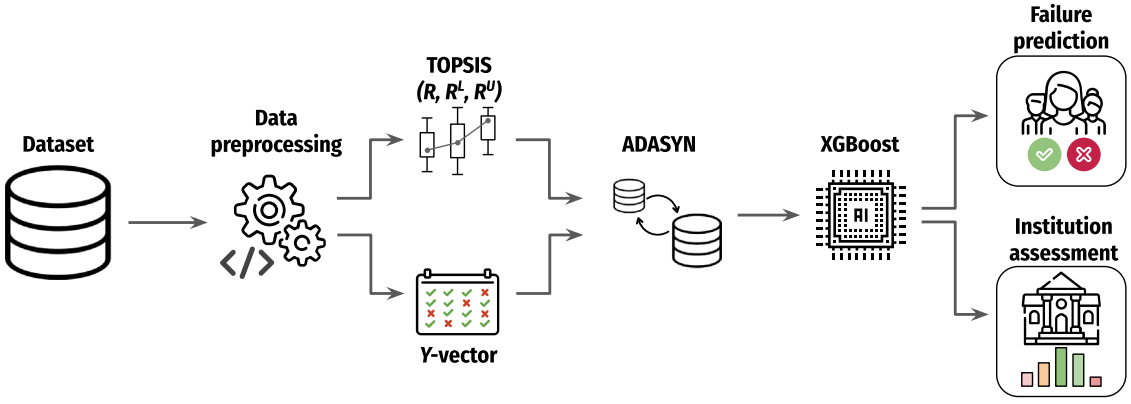

Fig. 2. Flowchart of the procedure conducted for generating our decision support system and its output.

4.3. MCDM feature extractor

In order to evaluate the initial scenario of each particular student, we have implemented both TOPSIS and uwTOPSIS techniques. On the one hand, the TOPSIS method has been applied because the weighting scheme has been meticulously chosen by the SEA-UIS [56]. This advantage allows us to understand how the institution notices the initial stage of people and then adds it to our dataset. On the other hand, the uwTOPSIS has allowed us to analyze the academic performance before entering college [12]. By applying a rank reversal approach, we can extract the (R_i^L, R_i^U)-pairs per each alternative and consider it as an additional feature that evaluates the a priori global situation of the set of students. The required ideals in step 3 of the unweighted algorithm have been fixed to avoid an output that depends on the entire dataset. In other words, we wanted to preserve their limitations in terms of educational attributes.

Despite having three different features (R_i, R_i^L, R_i^U) that might seem highly correlated, what we have merged with this process is an indicator of institution insight plus their relative boundaries in which it varies. Although many other features could be included in the model to improve the performance, we have decided to follow the guidelines of the managers of the UIS. In addition, the sensitivity of TOPSIS regarding the weighting scheme is widely studied [57], so the use of lower and upper-ranking intervals can attach robustness to the results. In turn, we have designed a dataset with size $N = 3534$ and $M = 8$ features, so each X_i array can be decomposed as $(A_i, C_i, E_i, H_i, S_i, R_i, R_i^L, R_i^U)$ to examine their impact on the classification task.

4.4. Model implementation

The implementation of the predictive system developed for this paper is illustrated in Fig. 2. It involves the pipeline of collection, preparation, normalization, feature extraction, model fitting, and output presentation. With respect to the output, we have indicated the two goals of this work, prediction of academic failure and assessment of the SEA-UIS.

For comparing and contrasting the results obtained by our classifier, we have decided to implement other classification tree-based methodologies to carry out an in-depth comparison. The models that have been trained with the Colombian students' dataset are XGBoost, Gradient boosting machine (GBM), Random Forests (RF), and Decision Trees (DT). To this end, we have performed three different learning procedures for testing the performance of our proposed system. First, the decision-tree models are trained over the original dataset, i.e. an imbalanced task. Second, the decision trees are trained with an ADASYN-oversampled dataset. Third, our predictive system is trained by employing hyperparameter tuning. The procedures are displayed in Fig. 3.

4.5. Evaluation metrics

The binary classification of students is a difficult task and it can be explained by means of various factors. First, the actual behavior of students is somehow chaotic because school dropout is a major concern in public universities. Second, the annotations associated with each student can be subject to uncertainties due to the number of different entities that handle such values. Third, we are just feeding our model with data stored by the university, therefore we are disregarding relevant historical information that concerns each individual.

Anyway, the data collection has been duly collected, annotated, and stored by the SEA-UIS, thus facilitating the goals of this paper. Nevertheless, we are facing a problem of imbalanced learning due to the number of students who pass the course being considerably greater than the ones who fail. In fact, it made us apply the ADASYN technique for oversampling. Considering a 2×2 confusion matrix, the metrics that will be used for model evaluation are the following equations:

$$\begin{aligned} &\text{Precision: } \frac{TP}{TP+FP}, \quad \text{Recall: } \frac{TP}{TP+FN}, \\ &F_1\text{-Score: } \frac{2 \cdot P \cdot R}{P+R}, \quad \text{Accuracy: } \frac{TP+TN}{TP+TN+FP+FP}. \end{aligned} \tag{17}$$

In order to carry out a more in-depth study, we have selected two additional metrics. On the one hand, Cohen's κ score for testing inter-rater reliability. Given a confusion matrix, the κ correlation statistic is computed as:

$$\kappa = 2\frac{TP \cdot TN - FN \cdot FP}{(TP+FP)\cdot(FP+TN)+(TP+FN)\cdot(FN+TN)}. \tag{18}$$

The κ score varies in $[-1, 1]$, indicating the percentage of agreement in the sample [58]. On the other, we compute the area under the receiver operating characteristic curve, which from now on is referred to as the AUC score. With this evaluation metric, we can control the misclassification errors that may occur and the error trade-offs [59].

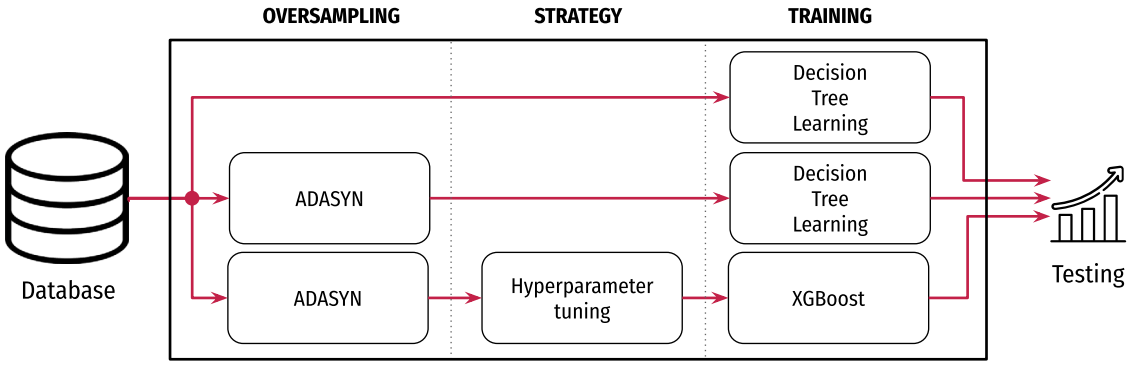

Fig. 3. Different learning procedures implemented for the student classification task.

Table 2
Grid with the hyperparameters involved during the cross-validation search and its optimal solution for the search.

Parameter	Grid values	Optimal value
D_{max}	5, 6, 7, 8, 9	7
DTs	$\{200 + 10i\}_{i=0}^{10}$	250
η	0.1, 0.05, 0.02, 0.01	0.1
γ	0.075, 0.1, 0.125, 0.15	0.1
λ	0.2, 0.25, 0.3, 0.5	0.25

4.6. Hyperparameter tuning

In order to get the best performance possible, we implemented a hyperparameter tuning algorithm. It is essential in most machine learning implementations because when more accurate, more applicable to further data. Since our problem is defined as an unbalanced task, we have considered the AUC score as the metric to maximize during this stage. With this aim, we have defined a computational grid for some of the basic parameters that XGBoost requires for its fitting. Two elements to take into account for boosting are the maximum depth of the trees (D_{max}) and the number of decision trees (DTs) to fix for our model. As we discussed in 3.4.2, Eq. (13) presents two parameters for regularization γ and λ. Finally, the learning rate (η) is always a value to select meticulously so as to balance the learning gain during training. In Table 2 the possible values have been presented for the cross-validation technique, which implies 3520 different combinations. The last column of Table 2 indicates the optimal solution of the discrete choice values.

With the exception of the DTs parameter, the selected values were affordably selected with a usual scheme of values. It is clear that the grid could have been extended to a high cardinality, however, the computational limitations have led us to limit the complexity of the problem. For the number of decision trees implemented, we empirically discover that with a value less than 200, we obtained an underfitting model, whilst with a value greater than 300, we got overfitting with no generalization techniques. Therefore, we designed a uniformly distributed sequence with differences of 10 units per step.

5. Results and discussion

In this section, we present the results obtained for the binary classification task of predicting academic failure. Since the goal of this paper is to solve two problems presented in the SEA, we have divided this section into two. Firstly, we study the performance of our intelligent decision support system and show the importance of each of its components. Secondly, we analyze the feature importance of the machine learning models used for our experiments from an XAI perspective in order to show the learned patterns.

5.1. Early prediction of students' failure

The sample of 3534 students has been serialized to get our experimental dataset $\{\mathbf{X}_i, y_i\}_i$, where each $\mathbf{X}_i \in \mathbb{R}^8$. For the training data split, we have decided to extract the 25% of the data (884 students) for the test set. Then, this number of elements that have not taken part in the training phase can be analyzed *a posteriori*, thus leading to more reliability experiments and ensuring its scalability. Table 3 contains the evaluation metrics selected in 4.5 with students that belong to the out-of-train set. In regard to the train-test partition for the synthetic dataset, we have split the 4968 samples considering the same proportion. Thus, the test set has 1242 samples for evaluating our classifiers.

When comparing the results in Table 3, we see that XGboost (XGB) is the best suited for the imbalanced environment, although we also note that random forest (RF) is the most precise and sensitive. It is worth mentioning that the application of ADASYN produces a significant improvement in every case. For the newly generated balanced set, we can see that XGBoost has better predictive performance as well. For the last experiment, considering the tuned configuration described in Table 2, the test results show that our intelligent system (XGB-T) outperforms all the last trained models. In particular, the resultant evaluation metrics are positioned beyond the last ones, indicating a better fitting when classifying, generalizing, and avoiding mismatching. Another fact that should be highlighted is that the κ coefficient is 0.5224, indicating a good index of agreement, and the AUC score is 0.8135, showing great performance.

In Fig. 4, we have displayed the confusion matrices per each of the classifiers in order to show the hits and the type I and II errors.

For a further assessment of the tuned classifier, we have displayed the rates of success–failure of the values of the confusion matrix as a way to represent and complement the evaluation metrics. These graphics are shown in Fig. 5.

5.2. Feature importance

The classification has been conducted through the sequential process of the input features $\mathbf{X}$. Once the training phase is done, we can study the impact of the model of the selected set of variables $(A_i, C_i, E_i, H_i, S_i, R_i, R_i^L, R_i^U)$. As we explained in Section 3.4.2, the XGBoost method has a tree-boosting ensemble architecture. Then we can extract the resultant assessment of each feature when classifying. The table embedded in Fig. 6 shows the importance of each feature for the trained models.

One of the most significant points is the relevance gains associated with the economic and health dimensions. For the economic variable, the minimal value achieved has changed from 0.0892 to 0.1689, which is almost a twice increment, and for the health one, the change was from 0.0798 to 0.1912, i.e. more than double. For the maximal values, we can note in the case of GBM that the combination of both means varied from 13.45% to 62.97%, which means an increase of its

Table 3
Evaluation of the student's classification task for the test set of the imbalanced (Original), ADASYN oversampling (ADASYN), and hyperparameter tuning strategies.

	Model	Precision	Recall	F_1-score	Accuracy	κ-score	AUC
Original	XGB	0.4118	0.0524	0.0930	0.6912	0.0266	0.6035
	GBM	0.3721	0.0599	0.1032	0.6855	0.0212	0.5994
	RF	0.4486	0.1798	0.2567	0.6855	0.1014	0.5714
	DT	0.4196	0.1760	0.2480	0.6776	0.0846	0.5534
ADASYN	XGB	0.8206	0.5639	0.6685	0.7080	0.4231	0.7677
	GBM	0.8205	0.5562	0.6630	0.7047	0.4170	0.7658
	RF	0.7863	0.5840	0.6702	0.6999	0.4057	0.7605
	DT	0.7987	0.5501	0.6515	0.6927	0.3929	0.7380
	XGB-T	0.8130	0.6793	0.7401	0.7611	0.5224	0.8135

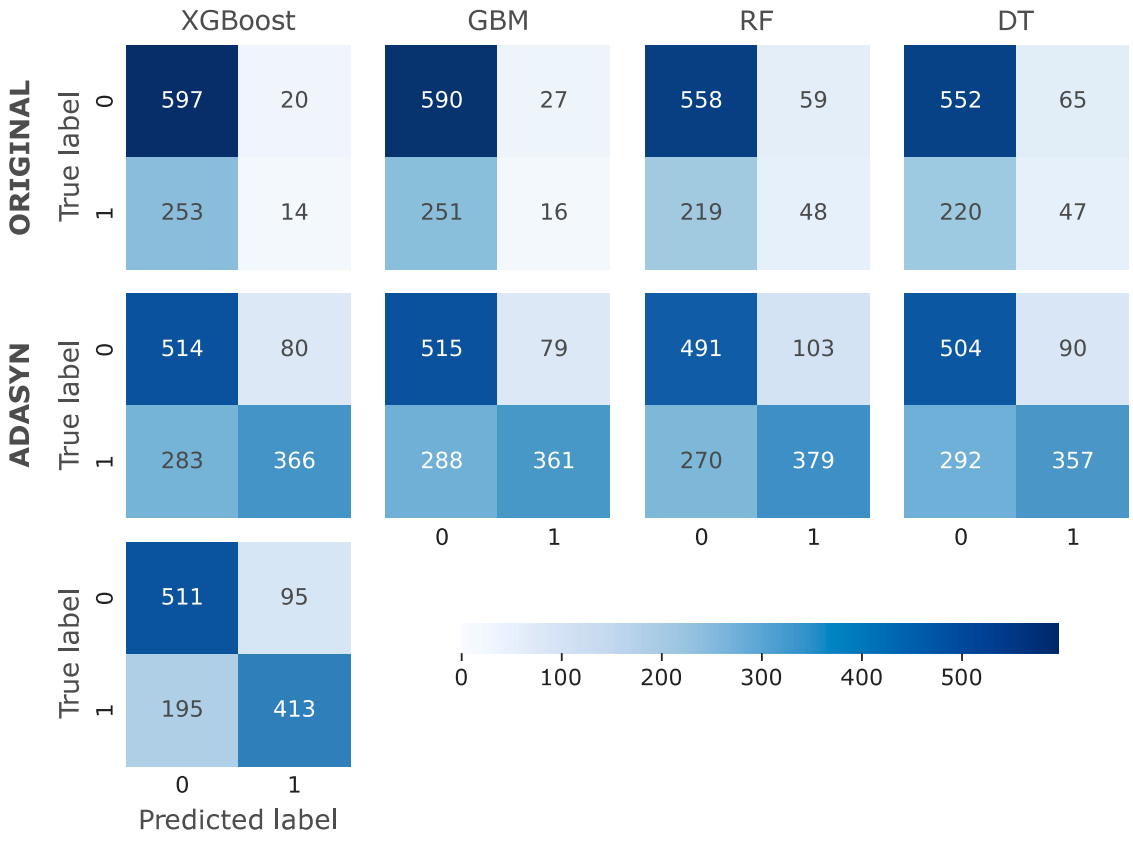

Fig. 4. Confusion matrices of each decision tree classifier for the imbalanced (first row), ADASYN (second row), and hyperparameter tuning (third row) strategies.

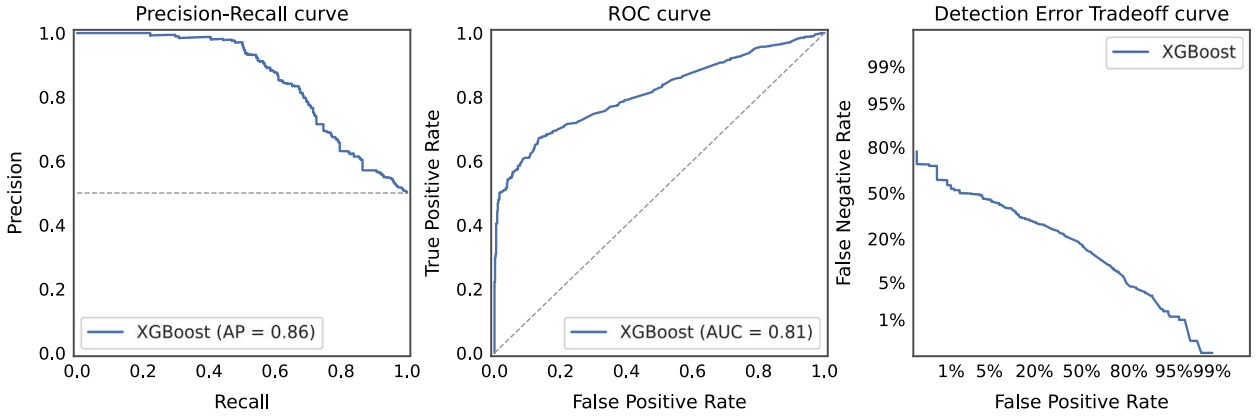

Fig. 5. Evaluation curves of the XGBoost after the hyperparameter tuning. From left to right, it is shown: First, the Precision–Recall Tradeoff for different probability thresholds with its average precision (AP). Second, the ROC curve with its AUC score (AUC). Third, the Type I and Type II error tradeoffs in percentage, also known as Detection Error Tradeoff (DET).

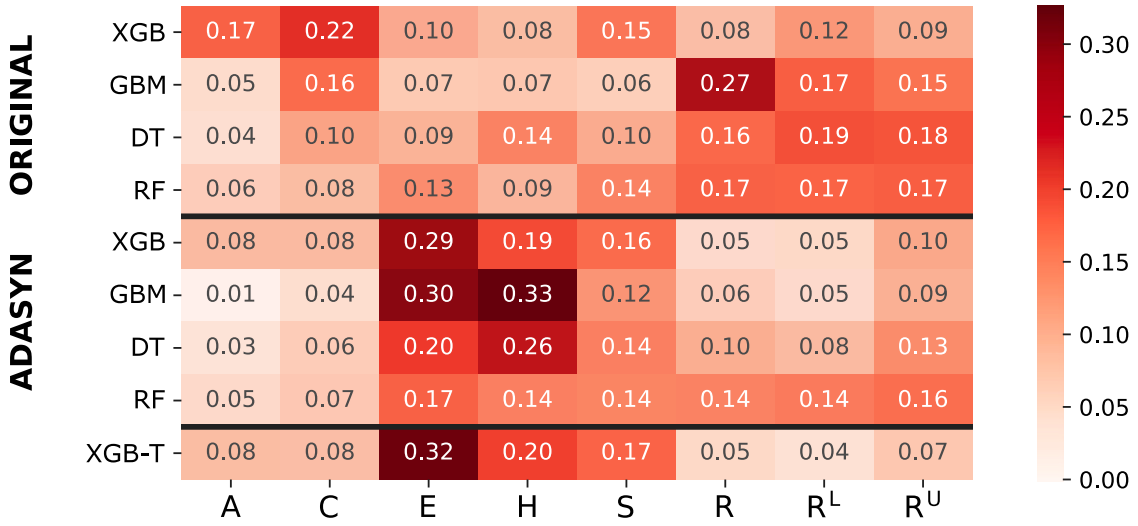

Fig. 6. Feature importance illustrated as a heat map of the model weights with values from 0 (dark color) to 1 (light color). Rows from 1 to 4 correspond to the results for the original dataset. Rows from 5 to 8 correspond to the results with the ADASYN technique. The last row is the combination of ADASYN plus hyperparameter tuning.

predictive power of 49.52 percentage units. We can also see a clear decreasing impact of the MCDM relative proximity values (R, R^L, R^U) when applying the synthetic oversample. Nevertheless, the sum of them means the 19.63% of the decision in its lowest case (XGBoost) and 44.01% in its greatest case (Random Forest).

It is interesting to check that before the oversampling technique was applied, the cognitive variable had a significant impact on the predictions. In cases such as XGBoost, more than the fifth part of the bundle depended on the C variable. On another note, the most stable feature is the social one because its variation has not been compromised by the ADASYN technique.

For the particular case of the optimized extreme gradient boosting model, it is necessary to emphasize that over half of the forecast impact (53.41%) is due to the economic-sanitary situation of the students before they entered the university. By adding the social dimension, it raises to 65.48%. Since none of them are directly correlated with the academic effort, it makes us think, once again, that marginal ditching is a major concern, because our model has succeeded in the 71.12% of the cases.

For a further visualization that illustrates the variation between feature importance, Fig. 6 shows a positional heat map of such values in a scale of $[0, 1]$. All the variations mentioned above have been reflected in such a chart. Due to the color mapping, we can see that the decision-weighted scheme varied from the (R, R^L, R^U) TOPSIS-array (imbalanced set) to the (E, H) features (synthetic set). As far as classifiers are concerned, Random Forest has been the model with minor modifications of their relative importance weights after the oversampling. It might explain why it has been the model with the best R and F_1 score in both approaches.

Regardless of the results obtained, we can say that the TOPSIS feature extraction has been very supportive of the classification problem. When we had no balance over our sample of students, the three values contributed to finding the correct response. Once we obtained a balanced case, even though their impact decreased, they still had a high impact on the decision-making. In short, their importance meant the 28.50% of the weighted system.

6. Conclusions

One of the main keys to academic success is the early detection of students' difficulties so that institutions can take action to correct them. With this in mind, the implementation of our XGBoost-based intelligent system makes it possible to estimate future results. As of this forecast, we can help managers of academic institutions to take measures and increase academic performance.

The predictive model utilized in this paper is powered by a tuned XGBoost algorithm. It has shown better results in comparison with other tree-based classifiers such as Gradient Boosting Machines, Random Forests, and Decision Trees. Moreover, the imbalanced problem attached to academic failure (or success) has been handled by applying the ADASYN oversampling technique. For the default Python implementation of the models, we have achieved a minimum AUC of 0.7380 for the decision tree case and a maximum of 0.7677 for the XGBoost case. This indicates the success of the tree-based algorithms for predicting academic failure. To conclude, we have implemented a cross-validated hyperparameter tuning for a reasonable grid of parameters which gave us an accuracy of 0.7611 and AUC of 0.8135. It was notable that the predictive power of our intelligent system is considerably good since the kappa coefficient achieved was 0.5224, showing great discrimination over the statistical type I and II errors.

The dataset used for this work belongs to the Universidad Industrial de Santander (UIS), a Colombian public university, where the managers have taken complementary action since the first course, aiming to reduce the educational failure of their students. We intend that, from our performance forecast, the institution can adapt its education policies and improve its results.

CRediT authorship contribution statement

Aarón López-García: Methodology, Software. **Olga Blasco-Blasco:** Conceptualization, Investigation, Review & editing. **Marina Liern-García:** Conceptualization, Writing, Resources. **Sandra E. Parada-Rico:** Data extraction, Conceptualization, Resources.

Declaration of competing interest

We confirm there is not actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations that could inappropriately influence, or be perceived to influence, their work.

Data availability

The data that has been used is confidential.

Acknowledgments

To the Academic Excellence Support System, Universidad Industrial de Santander (SEA-UIS), for the transfer of the information and data.

Funding

Sandra Parada-Rico would like to thank the Ministry of Science, Technology and Innovation, Colombia–MINCIENCIAS that is financing the research program code 1115-852 70767 and by the Ministry of Science and Technology code 70783, with resources from the "Autonomous heritage national fund for financing science, technology and innovation Francisco José de Caldas", contract CT 183-2021.

References

[1] WBG. World Bank Group: Research and development expenditure (% of GDP). 2020, https://datos.bancomundial.org/indicador/GB.XPD.RSDV.GD.ZS?most_recent_value_desc=true. [Accessed 07 September 2023].
[2] Jiménez MC. Abandono universitario tarea en la que Iberoamérica se sigue rajando. 2022, https://periodico.unal.edu.co/articulos/abandono-universitario-tarea-en-la-que-iberoamerica-se-sigue-rajando. [Accessed 20 April 2022].
[3] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining, vol. 13-17-Augu. 2016, p. 785–94.
[4] Freitas AA. Comprehensible classification models: A position paper. SIGKDD Explor Newsl 2014;15(1):1–10.
[5] Hwang C-L, Yoon K. Multiple attribute decision making: Methods and applications - a state-of-the-art survey. Lecture notes in economics and mathematical systems, Heidelberg: Springer Berlin; 1981.
[6] He H, Bai Y, Garcia EA, Li S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proceedings of the international joint conference on neural networks. 2008, p. 1322–8.
[7] Kabak M, Burmaoğlu S, Kazançoğlu Y. A fuzzy hybrid MCDM approach for professional selection. Expert Syst Appl 2012;39(3):3516–25.
[8] Pekkaya M. Career preference of university students: An application of MCDM methods. Procedia Econ Financ 2015;23:249–55, 2nd Global Conference on Business, Economics, Management and Tourism.
[9] Parada S, Blasco-Blasco O, Liern V. Construcción de indicadores basada en medidas de similitud con ideales. Una aplicación al cálculo de índices de adecuación y de excelencia.. Recta 2017;18:119–35.
[10] Parada SE, Blasco-Blasco O, Liern V. Adequacy indicators based on pre-established goals: An implementation in a Colombian University. Soc Indic Res 2019;143(1).
[11] Liern V, Parada-Rico S, Blasco-Blasco O. Construction of quality indicators based on pre-established goals: Application to a Colombian Public University. Mathematics 2020;8:1075.
[12] Blasco-Blasco O, Liern-García M, López-García A, Parada-Rico SE. An academic performance indicator using flexible multi-criteria methods. Mathematics 2021;9(19).
[13] Adams AJ, Hancock T. Work experience as a predictor of MBA performance. Coll Stud J 2000;34(2):211–7.
[14] Ganyaupfu EM. Teaching methods and students' academic performance. Int J Humanit Soc Sci Invent 2013;2(9):29–35.
[15] Mushtaq I, Khan SN. Factors affecting students' academic performance. Glob J Manag Bus Res 2012;12(9):17–22.
[16] Paliwal M, Kumar UA. A study of academic performance of business school graduates using neural network and statistical techniques. Expert Syst Appl 2009;36(4):7865–72.
[17] Imran M, Latif S, Mehmood D, Shah MS. Student academic performance prediction using supervised learning techniques. Int J Emerg Technol Learn 2019;14(14).
[18] Bhutto ES, Siddiqui IF, Arain QA, Anwar M. Predicting students' academic performance through supervised machine learning. In: 2020 international conference on information science and communication technology. 2020, p. 1–6.
[19] Verma U, Garg C, Bhushan M, Samant P, Kumar A, Negi A. Prediction of students' academic performance using machine learning techniques. In: 2022 international mobile and embedded technology conference. 2022, p. 151–6.
[20] Hidayah I, Permanasari AE, Ratwastuti N. Student classification for academic performance prediction using neuro fuzzy in a conventional classroom. In: 2013 International conference on information technology and electrical engineering. 2013, p. 221–5.
[21] Okubo F, Yamashita T, Shimada A, Ogata H. A neural network approach for students' performance prediction. In: Proceedings of the seventh international learning analytics & knowledge conference. New York, NY, USA: Association for Computing Machinery; 2017, p. 598–9.
[22] Sapiezynski P, Kassarnig V, Wilson C. Academic performance prediction in a gender-imbalanced environment. In: FATREC workshop on responsible recommendation proceedings. 2017, p. 49–58.
[23] Hasan R, Palaniappan S, Raziff ARA, Mahmood S, Sarker KU. Student academic performance prediction by using decision tree algorithm. In: 2018 4th International conference on computer and information sciences. 2018, p. 1–5.
[24] Awaji M. Evaluation of machine learning techniques for early identification of at-risk students (Ph.D. thesis), Nova Southeastern University; 2018.
[25] Keser SB, Aghalarova S. HELA: A novel hybrid ensemble learning algorithm for predicting academic performance of students. Educ Inf Technol 2022;27(4):4521–52.
[26] Mushava J, Murray M. A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. Expert Syst Appl 2022;202:117233.
[27] Thai-Nghe N, Busche A, Schmidt-Thieme L. Improving academic performance prediction by dealing with class imbalance. In: 2009 ninth international conference on intelligent systems design and applications. 2009, p. 878–83.
[28] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artificial Intelligence Res 2002;16(Sept. 28):321–57.
[29] Hussain S, Dahan NA, Ba-Alwib FM, Ribata N. Educational data mining and analysis of students' academic performance using WEKA. Indonesian J Electr Eng Comput Sci 2018;9(2):447–59.
[30] Delahoz-Dominguez E, Zuluaga R, Fontalvo-Herrera T. Dataset of academic performance evolution for engineering students. Data Brief 2020;30:105537.
[31] Cortez P. Student performance. 2014, http://dx.doi.org/10.24432/C5TG7T, UCI Machine Learning Repository.
[32] Loyola-González O. Black-box vs. White-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access 2019;7:154096–113.
[33] Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. J Eng Appl Sci 2017;12(16):4102–7.
[34] Fan C, Chen M, Wang X, Wang J, Huang B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Front Energy Res 2021;9.
[35] Luce RD, Raiffa H. Games and decisions: Introduction and critical survey. Philos Phenomenol Res 1958;19(1):122–3.
[36] Liern V, Pérez-Gladish B. Multiple criteria ranking method based on functional proximity index: Un-weighted TOPSIS. Ann Oper Res 2020;1–23.
[37] López-García A. uwTOPSIS. In: GitHub repository. 2021, GitHub, https://github.com/Aaron-AALG/uwTOPSIS.
[38] Deconinck E, Hancock T, Coomans D, Massart D, Heyden YV. Classification of drugs in absorption classes using the classification and regression trees (CART) methodology. J Pharm Biomed Anal 2005;39(1):91–103.
[39] Ho TK. Random decision forests. In: Proceedings of 3rd international conference on document analysis and recognition, vol. 1. 1995, p. 278–82.
[40] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189–232.
[41] Giannakas F, Troussas C, Krouska A, Sgouropoulou C, Voyiatzis I. Xgboost and deep neural network comparison: The case of teams' performance. In: Cristea AI, Troussas C, editors. Intelligent tutoring systems. Cham: Springer International Publishing; 2021, p. 343–9.
[42] Ogunleye A, Wang Q-G. XGBoost model for chronic kidney disease diagnosis. IEEE/ACM Trans Comput Biol Bioinform 2020;17(6):2131–40.
[43] Gumus M, Kiran MS. Crude oil price forecasting using xgboost. In: 2017 International conference on computer science and engineering. IEEE; 2017, p. 1100–3.
[44] Chang YC, Chang KH, Wu GJ. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. Appl Soft Comput 2018;73:914–20.
[45] He H, Zhang W, Zhang S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Syst Appl 2018;98:105–17.
[46] Dang LM, Hassan SI, Im S, Moon H. Face image manipulation detection based on a convolutional neural network. Expert Syst Appl 2019;129:156–68.
[47] Došilović FK, Brčić M, Hlupić N. Explainable artificial intelligence: A survey. In: 2018 41st international convention on information and communication technology, electronics and microelectronics. 2018, p. 0210–5.
[48] Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for trees: From local explanations to global understanding. 2019.
[49] Sagi O, Rokach L. Explainable decision forest: Transforming a decision forest into an interpretable tree. Inf Fusion 2020;61:124–38.
[50] Melo-Becerra LA, Ramos-Forero JE, Hernández-Santamaría PO. La educación superior en Colombia: situación actual y análisis de eficiencia. Revista Desarrollo Soc 2017;1(78):59–111.
[51] El Observatorio de la Universidad Colombiana, OUC. Educación Superior: Indicadores de Colombia, aún por debajo del promedio OCDE. 2023, https://www.universidad.edu.co/educacion-superior-indicadores-de-colombia-aun-por-debajo-del-promedio-ocde/. [Accessed 18 April 2023].
[52] SEA-UIS. Sistema de Excelencia Académica. Universidad Industrial Santander. 2022, https://www.uis.edu.co/webUIS/es/estudiantes/excelenciaAcademica/index.html. [Accessed 25 July 2022].

[53] Parada SE, Fiallo JE, Blasco-Blasco O. Construcción de indicadores sintéticos basados en Juicio experto: Aplicación a una medida integral de la excelencia académica. Recta 2015;16:51–67.
[54] Sola J, Sevilla J. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans Nucl Sci 1997;44(3):1464–8.
[55] Trebuňa P, Halčinová J, Fil'o M, Markovič J. The importance of normalization and standardization in the process of clustering. In: 2014 IEEE 12th international symposium on applied machine intelligence and informatics. 2014, p. 381–5.
[56] Blasco-Blasco O, Parada Rico SE, Liern-García M, López-García A. Characterization of university students through indicators of adequacy and excellence. Analysis from gender and socioeconomic status perspective. In: ICERI2020 proceedings. 13th annual international conference of education, research and innovation, IATED; 2020, p. 8030–7.
[57] Dutta B, Singha T, Goh M, Lamata M-T, Verdegay J-L. Post factum analysis in TOPSIS based decision making method. Expert Syst Appl 2019;138:112806.
[58] Wan T, Jun H, Zhang H, Pan W, Hua H. Kappa coefficient: a popular measure of rater agreement. Shanghai Arch Psychiatry 2015;27(1):62.
[59] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30(7):1145–59.