frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 02 October 2023
DOI 10.3389/feduc.2023.1244686

Predictive analytics study to determine undergraduate students at risk of dropout

Andres Gonzalez-Nucamendi, Julieta Noguez*, Luis Neri, Víctor Robledo-Rella and Rosa María Guadalupe García-Castelán

School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Mexico

Check for updates

OPEN ACCESS

EDITED BY
Julius Nganji,
University of Toronto, Canada

REVIEWED BY
Antonio Luque,
University of Almería, Spain
Najah Al-Shanableh,
Al al-Bayt University, Jordan

*CORRESPONDENCE
Julieta Noguez
✉ jnoguez@tec.mx

RECEIVED 22 June 2023
ACCEPTED 18 September 2023
PUBLISHED 02 October 2023

CITATION
Gonzalez-Nucamendi A, Noguez J, Neri L, Robledo-Rella V and García-Castelán RMG (2023) Predictive analytics study to determine undergraduate students at risk of dropout. *Front. Educ.* 8:1244686.
doi: 10.3389/feduc.2023.1244686

In this this work, a study is presented with quantitative variables using machine learning tools to detect undergraduate students at risk of dropping out and the factors associated with this behavior. Clustering algorithms and classification methods were tested to determine the predictive power of several variables regarding the dropout phenomenon on an unbalanced database of 14,495 undergraduate students with a real dropout rate of 8.5% and a retention rate of 91.5%. The usual classification criterion that assigns individuals to a class if their probability of belonging to it is greater than 50% provided accuracies of 13.2% in the dropout prediction and 99.4% in the retention prediction. Among eight classifiers, Random Forest was selected and applied along with Threshold Probability, which allowed us to gradually increase the dropout precision to more than 50%, while maintaining retention and global precisions above 70%. Through this study, it was found that the main variables associated with student dropouts were their academic performance during the early weeks of the first semester, their average grade in the previous academic levels, the previous mathematics score, and the entrance exam score. Other important variables were the number of class hours being taken, student age, funding status of scholarships, English level, and the number of dropped subjects in the early weeks. Given the trade-off between dropout and retention precisions, our results can guide educational institutions to focus on the most appropriate academic support strategies to help students at real risk of dropping out.

KEYWORDS

student dropout, learning analytics, predictive models, higher education, educational innovation, threshold probabilities, Artificial Intelligence

1. Introduction

The study of school dropouts is of interest at all educational levels. Reducing student dropouts is an important challenge that high schools and higher education must face. The loss of students who are beginning their high school or undergraduate studies constitutes a worldwide concern (e.g., Heublein, 2014; Aulck et al., 2016; Hsu and Yeh, 2019; Olaya et al., 2020). Several factors have been studied as the origins of dropping out, including unfavorable sociodemographic conditions, insufficient academic support, underprivileged economic income, and poor academic and social capabilities. Quantitative research on the causes and the possible solutions for dropping out has been reported in the literature (e.g., Aulck et al., 2016; Garg et al., 2021).

The difficulty in conducting extensive research on student dropouts is that many variables may play a simultaneously important role. For example, academic failure may be caused by a lack of clarity on the relevance of real-life scenarios (Cameron and Heckman, 2001; Wexler and Pyle, 2012) and a lack of motivation that gives rise to random class attendance. Moreover, some students may have high rates of behavior problems because of a weak family and home structure (Wexler and Pyle, 2012). Some families place unrealistic expectations on their daughters and sons without providing them with the required tools for success. The absence of committed parents and family academic role models may also play a key factor (Balfanz et al., 2007). On the other hand, first-generation students may need a supportive environment in their schools to compensate for the sometimes-non-existent academic structure found at home.

To address this problematic situation, the Institute of Education and Science (US) has provided six recommendations to prevent dropouts at school (Dynarski et al., 2008):

(1) Data systems should be aimed at the early detection of students at risk.
(2) One-on-one tutoring is highly recommended for this population.
(3) Academic support must be provided (extra office hours, extra homework, etc.).
(4) Teaching social skills and providing specific programs to round out the class experience should not be underestimated.
(5) Personalized academic instruction must be an option.
(6) Focusing on lifelong competencies in addition to rigorous and relevant instruction must also be considered.

Studying the impact of the diverse factors that produce dropping out in middle and higher education has pushed institutions to perform statistical studies to disclose the relative importance of these factors and to apply suitable and timely measures to predict students at risk of dropping out (Hsu and Yeh, 2019). In this regard, the incorporation of learning analytics techniques that involve simultaneous analysis of students' social and performance data can disclose the factors that have a larger impact on dropping out. These techniques have contributed to the improved accuracy of predictive models in recent years (e.g., Amare and Simonova, 2021; Saravanan et al., 2022). Nowadays, data analysis techniques are applied to large data sets to better understand the relationships among the multiple variables involved.

The present research should help improve the design of institutional retention programs by tailoring them to students who are at risk of dropping out but are the most likely to be retained according to appropriate selection algorithms. In this context, we use machine learning (ML) tools in terms of predictive analytics, to identify potential students at risk and define the characteristics that place them in such a situation. The information obtained will be useful in designing specific retention programs.

The objective of this paper is to find the most accurate predictive model that allows to make the best timely decisions for institutional intervention, considering its ability to predict relative percentages of students at risk of dropping out.

The research questions that guide the present study are:

(a) What are the main factors that cause undergraduate dropout?
(b) Which groups of students are the most vulnerable?

In this first phase of this work, a study is presented with quantitative variables using machine learning tools to detect undergraduate students at risk of dropping out and the factors associated with this behavior. The organization of the paper is as follows. In section "2. Theoretical framework," a theoretical framework regarding the use of ML and learning analytics to predict dropping out is presented. In section "3. Related work," related studies in the literature on student dropouts are briefly summarized. Section "4. Methodology" presents the methodology followed in the present research and the case study selected. Section "5. Results and analysis" includes the principal results and analysis. Section "6. Discussion" presents the discussion, and finally, in section "7. Conclusion and future work," the conclusions and future work are outlined.

2. Theoretical framework

The machine learning (ML) tools and concepts used in this research are briefly described below.

2.1. Machine learning tools

2.1.1. Grouping or clustering algorithms

Clustering algorithms are procedures for grouping a series of vectors, associated with the variables according to specific criteria. Those criteria are usually distance or similarity. The closeness between the vectors is defined with a selected distance function, such as the Euclidean, although other metrics may be used. Generally, vectors in the same group (or clusters) share common properties. The knowledge of the groups allows a synthetic description of a complex multidimensional data set (e.g., Romesburg, 2004).

There are two main techniques for grouping: (a) hierarchical grouping, which can be agglomerative or divisive, and (b) non-hierarchical grouping, in which the number of groups is determined in advance, and the observations are assigned to the groups based on their closeness. For the latter technique, there are *k*-means and *k*-medoids methods.

2.1.2. The *k*-means method

The *k*-means method is probably the most used when the data set is so large that the computational time of the Hierarchical Clustering method, which is undoubtedly more accurate, is too large. In *k*-means, the number of groups is selected *a priori* and randomly creates an equal number of centroids; therefore, *k*-means does not always generate the same assignments for different program runs with similar conditions.

2.1.3. Predictive power of variables

To visualize and analyze the predictive power of a specific numerical variable and distinguish between dropping out and retention, this work applied the technique of density functions. It yields a continuous function derived from smoothing a histogram of relative frequencies, so the area under the curve represents probabilities. The diagrams in **Figure 1** illustrate this mechanism.

When it comes to a categorical predictive variable with k categories, the predictive power can be visualized by fusing a bar chart to distinguish among the categories with a greater or lesser proportion of dropouts, as shown in Section 3.1.5 below.

2.1.4. Classification methods

Classification methods are used to assign individuals to specific groups based on previously defined characteristics. In our study, the main characteristics associated with student dropout were (a) their academic performance during the early weeks of the first semester, (b) their average grade in the previous academic levels, (c) the previous mathematics score, and (d) the entrance exam score. Other important variables were the number of class hours being taken, student age, funding status of scholarships, English level, and the number of dropped subjects in the early weeks of the academic period. Algorithms determine the combination of these characteristics that define an individual's membership in a category. Predictive models are machine learning techniques applied to databases that seek to identify patterns to predict the membership of individuals in categories and make informed decisions. The predictive area has recently assumed a leading role in education (e.g., Liu et al., 2022).

We have selected eight classifiers from a wide range of available options, based on our previous experience and the diversity of approaches they offer. These classifiers were chosen specifically to address our classification problem. The list includes Support Vector Machine (SVM), which searches for a separating hyperplane in a feature space (Cortes and Vapnik, 1995); K-Nearest Neighbors (KNN), which classifies based on closeness to the nearest K data points (Cover and Hart, 1967); Decision Trees, a hierarchical structure that makes classification or regression decisions using nodes representing feature questions (Quinlan, 1986); Random Forest, an ensemble of decision trees that combines results to improve accuracy (Breiman, 2001); ADA Boosting (Adaptive Boosting), an ensemble that improves weak classifiers by assigning greater weight to incorrectly classified instances (Freund and Schapire, 1996); Extreme Gradient Boosting (XGBoost), an efficient implementation of boosting with multiple decision trees (Chen and Guestrin, 2016); Naive Bayes, a probabilistic classifier based on Bayes' theorem that assumes independence among features (Duda et al., 2001); and LDA (Linear Discriminant Analysis), which finds linear combinations of features for discrimination between classes (Fisher, 1936). These classifiers were selected for their versatility and ability to address a wide variety of approaches to solving our problem.

The evaluation of a predictive model is based on a confusion matrix, which is a valuable tool to assess how well an ML classification model works. It is used to show explicitly when one class is confused with another, which allows working separately with different error measures. *Positive precision* refers to the dropout cases and *negative precision* refers to the retention cases. Therefore, these values and the *overall accuracy* of the prediction can be obtained as follows:

Positive precision: Percentage correctly classified as dropout;

$$PP = TP/(TP + FP).$$

Negative precision: Percentage correctly classified as retention;

$$NP = TN/(FN + TN).$$

Overall accuracy: Percentage the total number of cases correctly classified:

$$OA = (TP + TN)/(TP + TN + FP + FN).$$

Although F-Measure is commonly used to compare classifiers (Powers, 2020) it has limitations in situations of unbalanced classes and varying probability thresholds. This is because it varies by threshold, which makes comparisons difficult. In addition, it is sensitive to class imbalance and may bias the evaluation toward the majority class. For these reasons, we chose to visually assess the performance of the eight classifiers through a scatter plot (**Figure 2**) showing the probability of dropout on the X-axis and the probability of retention on the Y-axis. We observe that Random Forest, with high "accuracy," stands out as a promising option that requires less tuning to achieve good performance. This leads us to prefer Random Forest over other classifiers that demand exhaustive hyperparameter settings.

Additionally, Random Forest is a technique of great importance in the analysis of dropout data since it allows us to visualize the importance of the predictor variables. We can obtain a graph showing the relative importance of the variables and their individual effect on model improvement, i.e., how much the overall accuracy of the model is damaged by considering the absence of each variable in the whole forest. The most important variable is assigned 100%, and the others are given relative importance in the form of a number between 0 and 1. Then, a standardization is performed so that the sum of all contributions equals 100%.

2.1.5. Threshold probability method as assignation criterium

When the variable to be predicted is highly unbalanced, as is the case for the retention and dropout cases in the example shown in **Figure 3**, a bias toward the dominant class may occur, even when classifying all individuals in that class. This usually happens when using a fixed probability threshold. For example, we can consider that a variable to predict A, has two categories: Yes or No. Traditionally, A = Yes is assigned whenever the probability $P(A = \text{Yes}) > 0.5$; and A = No, is assigned otherwise. This logic works well when the training data is balanced, that is, when it contains approximately the same number of Yes and No cases. However, this does not happen in real scenarios where there is a large imbalance. For example, suppose that a database has only 10% of Yes cases and 90% of No cases. Under these circumstances, the overall precision measure is misleading because a naïve (and useless) rule assigning all cases to No would have a global precision of 90%, with a precision of No at 100% but a precision of Yes at 0%. Generally, the accuracy of Yes is the most interesting in real cases and therefore, in the given example, this assignation would be useless.

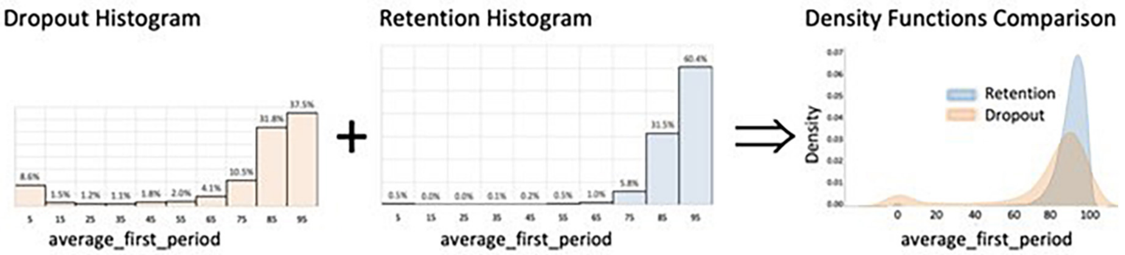

FIGURE 1
Example illustrating density functions.

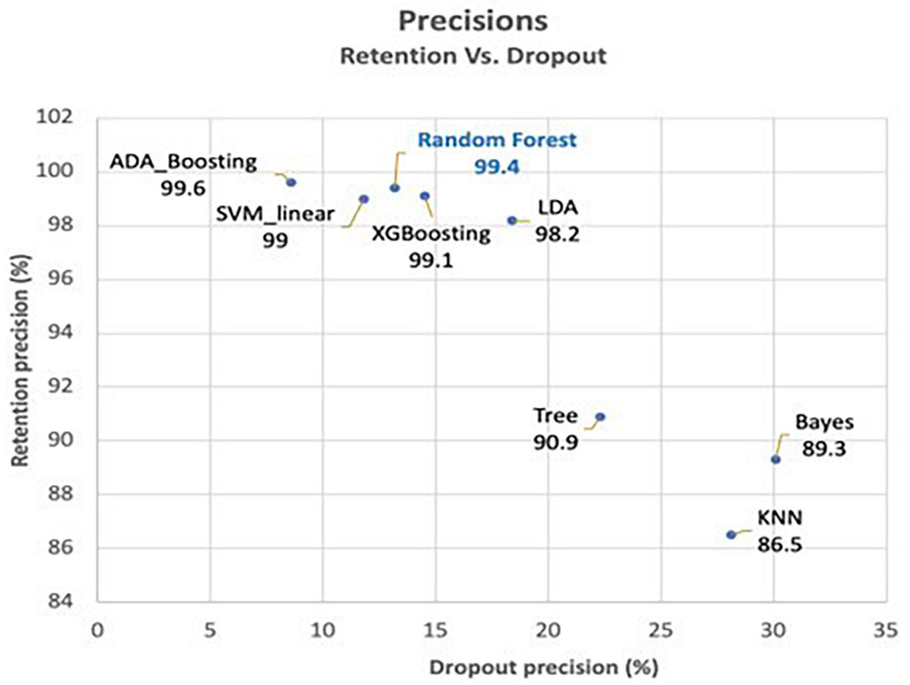

FIGURE 2
Comparison of retention and dropout precision classification methods. The retention percentage is indicated.

One way to reduce this problem is by varying the *cut-off probabilities*, requiring more probability from the class with the larger number of cases, and consequently, requiring less probability from the class with fewer cases. For example, we could manually set the rule $P(A = \text{Yes}) > = 0.2$ to assign Yes and consequently $P(A = \text{No}) > 0.8$ to assign No, and a better balance for the prediction probabilities for both classes would be achieved. Nevertheless, it is not recommended to assign such a low threshold chance to the non-dominant class to attain an accuracy closer to 100%, since it would be at the cost of huge damage to the dominant class accuracy (see Section "5.3.3. Predictive power of explicative variables" below).

2.2. Dropouts

Dropping out is a situation in which the student withdraws from an educational institution or system without obtaining accreditation or a school certificate (e.g., Lamb et al., 2010). It can occur at any educational level and is a complex problem due to many endogenous and exogenous variables, as presented in section "4. Methodology." Endogenous variables refer to the intrinsic characteristics of students, for example, their ability to learn, their interest in school, or their level of development. Exogenous variables are related to external factors such as economic factors, family conditions, and natural disasters.

3. Related work

Dropout models deal with complex issues in which individual choices, institutional processes, demographic background, health issues, teachers' opinions, student behavior and social factors play a role when a student decide to whether or not remain at the University (Hedge and Prageeth, 2018). The inability to cope with the performance demands of higher education institutions, wrong expectations, financial problems, keeping pace with lecturers, rowdy classrooms, time management and less identification with the career path are the most important reasons for dropping out (Aulck et al., 2016; Govender, 2020). GPAs in math, English, chemistry, and psychology, as well as birth year were among the strongest predictors of student persistence (Aulck et al., 2016). Student records and transcripts for courses taught in the first 2 years, high school averages, and whether the student graduated from the chosen major or not are all valuable input variables in the dropping out understanding (Abu-Oda and El-Halees, 2015; Von Hippel and Hofflinger, 2020). High school performance in humanities has a surprisingly significant impact even on engineering students (Nagy and Molontay, 2018).

Considering all the mentioned issues, Germany has considered academic policies that include broad assistance measures, such as flexibilization of the curricula, better information for students, and the expansion of support offered at the start of their studies (Heublein, 2014). Palestinian studies found out that digital design and algorithm analysis have a great effect on predicting student persistence in the major and decreasing the likelihood of students dropping out (Abu-Oda and El-Halees, 2015).

Some other universities around the world have used Machine Learning techniques, Naive-Bayes Classification Algorithms programmed in R, Gradient Boosted Trees, Deep Learning, rough set theory and k-means in an effort to determine the factors that influence dropping out (e.g., Abu-Oda and El-Halees, 2015; Aulck et al., 2016; Hedge and Prageeth, 2018; Nagy and Molontay, 2018; Olaya et al., 2020; Von Hippel and Hofflinger, 2020). Even Thematic analysis has been used to analyze qualitative narrative data (e.g., Govender, 2020).

Table 1 shows a comparison of the related work outlined above. In column 2 the statistical technique used, or the approach followed, by the different authors are outlined. In columns 3 to 10 the most relevant dropout factors extracted from among these references are indicated. The "x" signs indicate the dropout factors considered in each reference. From this table it is seen that the three most common and relevant factors for dropping out are: (a) first-year undergraduate grades, (b) high school grades, and (c) university entrance exam scores.

The following section describes the research methodology used in this study.

4. Methodology

To build predictive models to identify high-risk students in a timely manner we followed the research methodology indicated in **Figure 4**: (a) Case study selection: (b) Data cleansing and definition of the study variables; (c) Identification of relevant database subsets; (d) Definition of the research hypothesis; (e) Application of statistical and ML techniques; (f) Results and analysis; (g) Discussion; and (h) Conclusions.

4.1. Case study

As a case study, the analysis of dropout cases between 2014 and 2021 for a prominent private university in Mexico was chosen. Approximately 8.1% of the students who entered this institution did not manage to finish their studies or transferred to other institutions (Alvarado-Uribe et al., 2022). Although this dropout rate is low compared to the average for other Mexican universities, it does represent an important social cost and economic effort for families to support their sons' and daughters' studies. Moreover, school fees may be absorbed not only by parents or families but also by other institutions that regularly provide scholarships. In the second phase of this work, we are expanding the study to include qualitative variables such as socio-economic categories and social lag, that will be reported in a future work.

4.2. Data cleansing

We analyzed an institutional initial database of 143,326 records (students) with 50 independent variables (Alvarado-Uribe et al., 2022). To proceed with this research, a careful study and cleansing of the initial database yielded a suitable database for applying the selected ML techniques.

4.3. Sample breakdown

The initial database contained data from 2014 to 2022 including high school and college students. The institution launched a new educational model at the undergraduate level (Tec21, 2022) in the fall of 2019 (August–December 2019), so this research focuses on first-year undergraduate students enrolled in this new educational model to determine the variables that most influence dropouts and to propose intervention schemes.

4.4. Hypothesis

Derived from the research questions, the following hypothesis was established:

1. *It is possible to identify in a timely manner the key differentiating characteristics of undergraduate dropouts, and to cluster students for timely and adequate support.*

4.5. Machine learning analysis strategies

The ML analysis strategies considered were: (a) clustering; (b) classification methods comparing populations of dropouts and non-dropouts, where eight classification techniques were considered; (c) Random Forest in detail and Threshold Probability Method (TPM); and (d) the predictive power of the variables. The following section shows and analyzes the results.

5. Results and analysis

According to the established methodology (**Figure 4**), the most important results of each step are described below.

TABLE 1 Comparison of related work.

References	Technique	University first-year academic records	Entrance exam scores	High School grades	First choice major denied	Unfavorable sociodemograp conditions	Wrong expectations	Underprivileged economic income	Insufficient academic support
Relevant dropout factors									
Heublein, 2014	Empirical research	×	×	×	×	×	×	×	×
Abu-Oda and El-Halees, 2015	Decision Tree, Naive-Bayes, *k*-means, linear models, deep learning	×	×	×	×				
Aulck et al., 2016	Regularized logistic regression, *k* nearest neighbors, random forest	×				×	×	×	×
Hedge and Prageeth, 2018	Decision tree, Naive-Bayes, *k*-means, linear models, deep learning	×	×	×	×				
Nagy and Molontay, 2018	Decision tree-based algorithms, Naive Bayes, k-NN, linear models, and deep learning			×					
Hsu and Yeh, 2019	Hybrid approach: *k*-means, set theory	×	×	×					
Olaya et al., 2020	Uplift modeling	×	×	×	×	×	×	×	×
Von Hippel and Hofflinger, 2020	Simple logistic regression	×	×	×	×				
Govender, 2020	Thematic analysis	×					×		×

TABLE 2 Cluster characteristics.

Cluster number	Cluster size (students)	Retention (students)	Dropouts (%)
Cluster 0	8,230	7,713	6.3
Cluster 1	743	621	16.4
Cluster 2	5,522	4,927	10.8

5.1. Database description

As mentioned above, a curated database (DB) provided by our Institution was used as the data source (Alvarado-Uribe et al., 2022). This initial database included N_{tot} = 143,326 students and 50 academic/demographic variables. The DB contained information on 65,809 high school students and 77,517 undergraduate students enrolled from August–December 2014 to August–December 2020.

Supplementary Appendix Table 1 presents the 16 variables selected in the first phase of the study from the initial database. The variable name, their description, and their type are specified. Numerical variables use a continuous range of values within a given numerical interval, while categorical variables use a discrete set of data. The 16 variables used in this study include 14 numerical and 2 categorical.

5.2. Undergraduate student sample

The following analysis is divided into two parts: (a) a description of the undergraduate students' sample and the cleansing process, and (b) the derived results applying different ML analysis strategies to the cleaned sample. Both the data cleansing and the algorithm execution were programmed in Python using *NumPy, Pandas, Matplotlib*, and *Scikit-learn* libraries.

5.2.1. Cleansed used variables

Although the original sample consisted of 143,326 students, this research focused only on the 77,517 students in the undergraduate sample. From the undergraduate subset, only the 24,507 first-year students enrolled in the educational model (Tec21, 2022) at the Institution were considered. However, when making the selection of the 16 numerical variables, it was identified that many students did not have defined values for these variables, so it was necessary to eliminate those students from the sample. The homogeneous sample without empty entries considered 14,495 complete records. This is the final cleansed sample to which the machine learning analysis strategies described below were applied. It is important to state that the variable to be predicted in this research is the *retention* variable (number 16 in **Supplementary Appendix Table 1**).

5.3. Machine learning analysis strategies

The ML analysis strategies comparing dropout and non-dropout populations are: (a) clustering, (b) classification methods, (c) Random Forest in detail with Threshold Probability Method (TPM), which is helpful for unbalanced data classification (Rodríguez Rojas, 2022), and (d) predictive power of the variables.

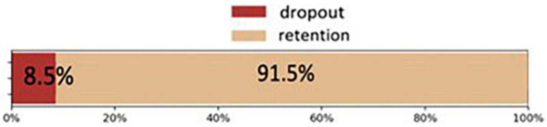

FIGURE 3
Example of unbalanced dropout and retention categories, taken from the study case selected (Section "4.1. Case study").

In (b), eight classification techniques were considered: (1) Support Vector Machine (SVM), (2) *k*-Nearest Neighbor (KNN), (3) Decision Trees (DT), (4) Random Forest (RF), (5) Adaptive Boosting (ADA_Boosting), (6) Extreme Gradient (XG_Boosting), (7) Bayesian Classifier (BC), and (8) Linear Discriminant Analysis (LDA). Below are the main results.

5.3.1. *k*-means clustering

To explore the prediction through cluster formation, the cleansed database was segmented into three main clusters, using the *k*-means technique. Cluster 0 had 8,230 students, 7,713 were retained and 517 dropped out. Cluster 1 had 743 students, 621 were retained and 122 dropped out. Cluster 2 had 5,522 students, 4,927 were retained and 595 dropped out. The corresponding sizes, retention numbers, and dropout percentages are shown in **Table 2**. The characteristics of the clusters are analyzed below.

In **Figure 5**, the cluster sizes are presented in a pie chart, and in **Figure 6** the corresponding dropout percentages are shown in a bar chart. In **Figure 7**, a radar diagram emphasizes the main variables related to student dropouts for each cluster. The radar diagram is normalized to amplify the differences among clusters, giving values between 0 and 100% to those variables with the lowest and largest weights, respectively. **Figure 7** presents the 14 explicative variables and the *dropout_semester* variable.

It should be noted that the information provided by the *dropout_semester* variable is equivalent to that of the retention variable. In fact, the value *dropout_semester* = 0 is equivalent to retention, and values *dropout_semester* = 1, 2, 3, or 4 are equivalent to dropout, that is, the *dropout_semester* variable is the breakdown in semesters of the dichotomous variable to be predicted, retention. Therefore, **Figure 7** shows the relative weight among clusters of the 14 explicative variables selected in this research to explain the variable to be predicted (*dropout_semest*er or *retention*).

Figure 6 shows that Cluster 0, representing 38% of the student sample, has the lowest dropout percentage, 6.3%. This cluster is characterized by students who (see **Figure 7**) have: (a) an intermediate percentage of dropped subjects or failed subjects during the first period of the first semester, (b) the highest average grade in the first period of the first semester, (c) the highest percentages of scholarship and loans, (d) the highest percentage of full-time students, (e) the highest general math evaluation, admission rubric score, and English evaluation, (f) the lowest percentage of students who took the admission test online, (g) the highest admission test and previous-level average-scores, and (h) the youngest students of the sample.

Cluster 1 is the smallest (5% of the student sample) and has the highest dropout percentage (16.4%). This cluster is characterized by: (a) the oldest students in their class, (b) the highest percentage

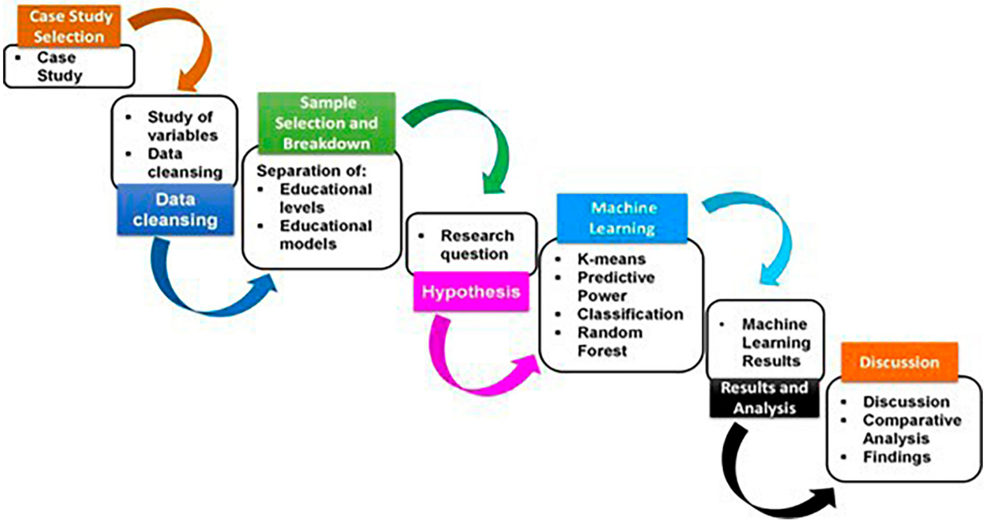

FIGURE 4
Steps of the research.

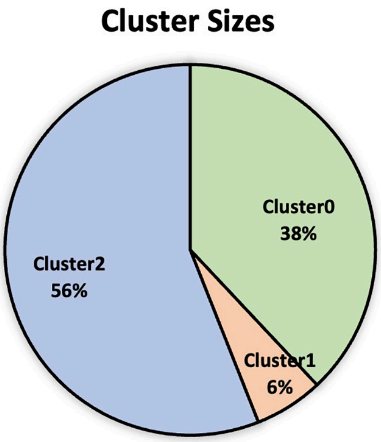

FIGURE 5
The cluster sizes.

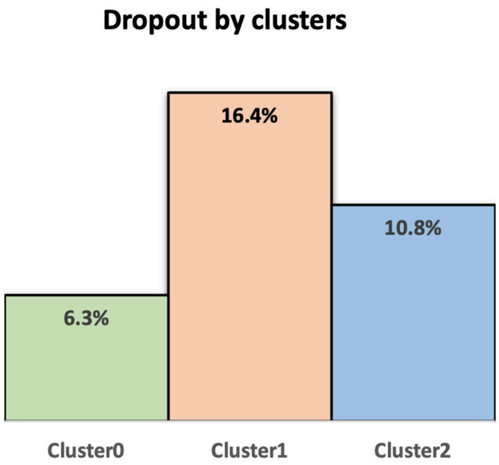

FIGURE 6
Dropout percentages by cluster.

of dropouts during the first year (*dropout_semester*), (c) relatively low percentages of dropped subjects or failed subjects during the first period of the first semester, (d) a low average grade in the first period of the first semester, (e) low percentage of scholarship or loans, (f) low percentage of full-time students, (g) a very low general math evaluation, admission rubric score, and English evaluation, (h) the highest percentage of students taking the admission test online, and (i) the lowest admission test and lowest previous level average scores.

Finally, Cluster 2 represents 57% of the student sample and has a relatively high 10.8% dropout percentage. This cluster is characterized by students: (a) with a high percentage of dropped subjects and failed subjects in the first period of the first semester, (b) the lowest average grade during the first period of the first semester, (c) no significant scholarships or student loans, (d) an intermediate percentage of full-time students, (e) general math evaluation, admission rubric and English evaluation that were average within the student sample, (f) a low percentage taking the admission test online, (g) admission test and previous-level average-score at an intermediate level, and (h) an age between those of clusters 0 and 1.

5.3.2. Classification methods

Several ML classifiers were tested to obtain the best accuracy for dropping out, retention, and/or global percentages. Through the execution of by-default parameters that required just a few adjustments, the classifiers gave dropout, retention, and global precision percentages for the undergraduate students in the sample ($N = 14{,}495$). The results are shown in Table 3. The comparison of the precisions obtained for the retention and dropout percentages by each classifier is shown in Figure 2 above.

The graph shows a negative relationship between the percentages of dropout precision and the percentage of retention

TABLE 3 Comparison of classifiers with by-default parameters.

Classifiers	Precision percentage (%)		
	Global	Dropout	Retention
SVM_linear	91	11.8	99.0
KNN	82	28.1	86.5
Decision Tree	85	22.3	90.9
Random Forest	92	13.2	99.4
ADA_Boosting	92	8.6	99.6
XGBoosting	92	14.5	99.1
Bayes	84	30.1	89.3
LDA	91	18.4	98.2

TABLE 4 Random Forest Dropout, Retention, and Global prediction precisions for different threshold probabilities.

Threshold probability	Dropouts	Retention	Global
0.000	0.000	1.000	0.915
0.400	0.060	0.997	0.917
0.500	0.120	0.995	0.920
0.600	0.181	0.990	0.920
0.700	0.233	0.983	0.918
0.750	0.257	0.970	0.909
0.880	0.510	0.811	0.785
0.889	0.534	0.783	0.761
0.901	0.586	0.745	0.731
0.945	0.755	0.480	0.504
0.995	1.000	0.019	0.103
1.000	1.000	0.000	0.086

precision. Overall, increasing the precision in dropout percentage yields a decreasing precision in the retention percentage. This fact can be used to obtain the optimum combination according to the requirements of each educational institution. There are only two exceptions to this rule. The first is in which RF (Random Forest) produces both higher dropout and retention precisions than SVM. The second is found in the lowest part of the retention precisions, where Bayes dominates over KNN, with higher precisions both in dropping out and retention. Notice that in the graph, all percentages of retention precision are greater than those of dropout precision. This is due to the imbalance in the variable that will be predicted (dropout percentage vs. retention percentage: 8.5% vs. 91.5%, respectively) as can be seen in **Figure 3** above.

5.3.3. Random forest and threshold probability method (TPM)

Using the Random Forest (RF) classifier with *n_estimators* = 400, criterion = *gini, min_samples_split* = 18, and changing the threshold probabilities, a threshold-probability graph was obtained. The results are shown in **Table 4**. **Figure 8** plots the graphs of the dropout precision, retention precision, and global precision of the sample. Threshold probability is along the horizontal axis and the vertical axis corresponds to the precisions (dropout, retention, and global). The training sample contained 80% of the cases and the remaining 20% corresponded to the testing sample. To obtain meaningful values for the comparison, the random seed was fixed as *random_state* = 0.

TABLE 5 Confusion matrix used to calculate effectiveness coefficients.

		Dropout	Retention	Real totals
Prediction				
Actual values	Dropout	127	122	249
	Retention	501	2,149	2,650
	Predicted totals	628	2,271	2,899

For an educational institution, it is possible to take advantage of these results to plan how to distribute the resources in retention efforts. For example, if the intervention point is selected at 51% dropout precision and 81% for retention according to **Table 4** (or **Figure 8**) the threshold probability is 0.88, and the global precision for the total sample is 78.5%. A measure that can be obtained with these results is the *effectiveness coefficient*, defined as the expected number of effective interventions that the institution should offer to students correctly predicted as dropouts divided by the total number of interventions the institution would offer to any student predicted as dropout (correctly or incorrectly) according to the model. To explain this coefficient, we use the testing sample that results from the remaining 20% of the students that were not included in the training sample. The total number of records in the testing table is therefore 0.2 × 14,495 = 2,899. The corresponding confusion matrix is presented in **Table 5**.

Out of the 249 real dropouts in the database for the testing sample, 127 were correctly detected and addressed, but 122 were undetected and, consequently, left unattended. Therefore, the precision in dropouts is $127/249 \approx 51\%$. Similarly, out of the 628 students predicted as dropouts, only 127 were true dropouts and 501 were false dropouts. The expected effectiveness is then $127/628 \approx 20\%$. If academic institutions implement intervention programs to attend to this population at risk of dropping out, only 1 out of 5 students will need these programs, while the remaining 4 will not, leading to misspending of valuable academic and economical resources. Nevertheless, note that for a random intervention the effectiveness would still decrease to only 8.5%, which is the total dropout percentage for the entire $N = 14{,}495$ undergraduate student sample, representing an even greater waste of academic and economic resources. Therefore, due to the imbalance among the types of variables to be predicted, it is useful to apply the threshold-probability method, because it can vary the precisions in the prediction of the dropout and of the retention class. This can guide institutions to implement the best interventions to address dropout cases.

5.3.4. Random forest important variables

Different tests were applied with several classifiers. Random Forest (RF) was chosen because of its good performance in retention and dropout percentages precisions, and because it

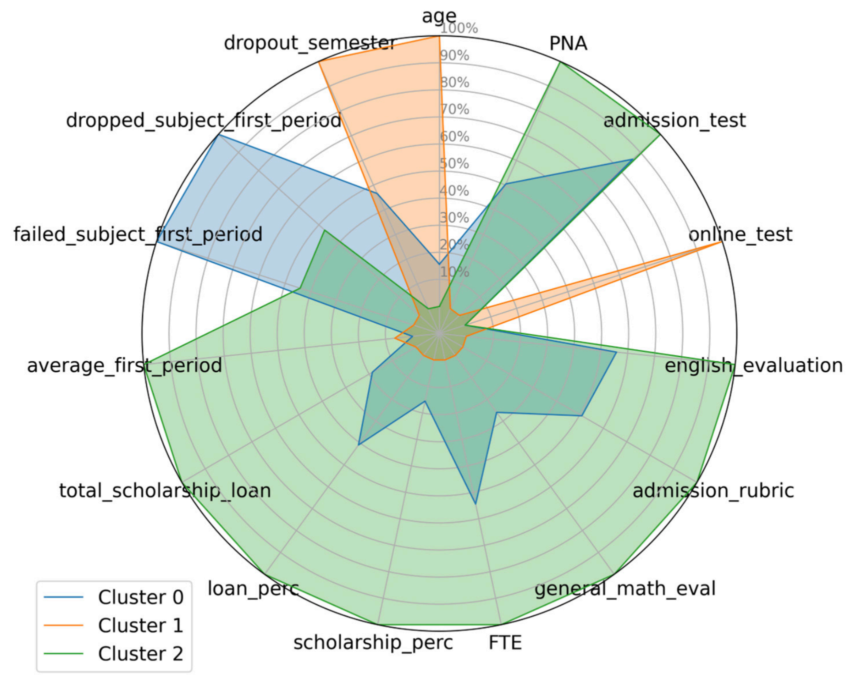

FIGURE 7
Cluster comparison for the selected variables (see **Supplementary Appendix Table 1** for variable names).

provided information on the importance of the used variables. **Figure 9** shows a sketch with the relative importance of the variables used, according to RF. It is important to notice here that only 11 of the 14 explicative variables were selected in RF, excluding those that complicated or even damaged the precision of the classifier. It is seen that, according to the RF classifier, the most important variable associated with student dropout is the average grade obtained in the first period of the first semester. Other important variables are: (a) the previous level average score (PNA), (b) the results of the general math evaluation (of the admission test and/or from the school of origin), (c) the admission test and admission rubric results, (d) full-time student status (FTE), (e) the student's age (younger students have lower dropout percentages than older ones, as mentioned in section "5.3. Machine learning analysis strategies"), (f) the total scholarships and student loans, and (g) the English evaluation result. According to RF, less important variables are the number of dropped out subjects in their first period or if the student took the admission test online.

5.3.5. Predictive power of explicative variables

To better know the influence of each variable in the dropout prediction, density function comparisons were made for the dropout class and the retention class, as explained in section "2.1.3. Predictive power of variables" (see **Figure 1**). The diagrams are presented in **Figure 10**, in order of importance according to the RF classifier. These diagrams reinforce the results already presented in **Figure 9**. The horizontal axis represents the range of possible values of the variables and the vertical axis shows the relative importance of that variable for predicting retention (in blue) and dropout (in peach color) cases. For example, from **Figure 10.1**, if students obtain in their first period a grade higher than 90, they will most probably be retained. On the other hand, if their grade was lower than 80, they were more likely to drop out. Similarly, **Figure 10.2** suggests that students with previous-level average grades (PNA) higher than 90 will likely be retained, while students with previous-level average grades below 80 are more likely to drop out. Similar conclusions can be seen in **Figures 10.3–10.12**. Bumps in the *x*-axis may correspond to input variables discretization.

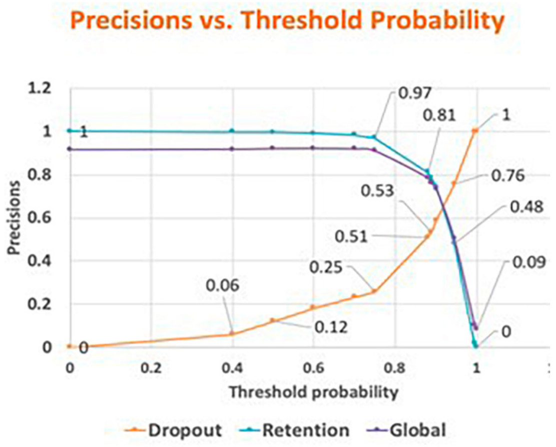

FIGURE 8
Random Forest Dropout, Retention and Global precisions vs. threshold probability.

6. Discussion

There are several methods to address the problem of class imbalance in the context of machine learning (Douzas et al., 2018), which refers to the situation where one of the classes

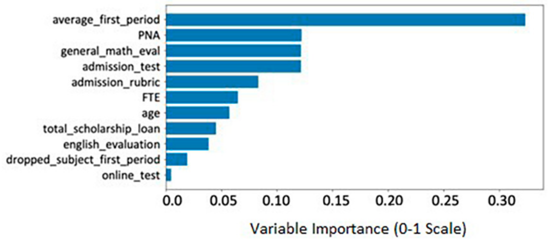

FIGURE 9
Important variables resulting from the application of Random Forest.

in a dataset is significantly smaller than the other. Many of these methods involve generating synthetic data, either through under sampling (removing records from the majority class) or oversampling (creating records in the minority class). However, this strategy raises the concern that it may distort the reality of the data, which in turn could affect the accuracy of algorithms when faced with new data.

An important observation is that, to date, we have not found references in the literature that use the probability threshold method in the context of student dropout. This suggests a scarcity in the application of this method in this particular area. Therefore, one of the main contributions of this article lies in presenting the probability threshold approach to address the problem of predicting student dropout.

The probability threshold approach involves varying the cutoff probability in the assignment criterion to either class. In the context of student dropout, this method offers an additional advantage, as it allows for the modulation (or selection) of the balance between “Yes” rates (Recall) and “No” rates (Sensitivity) according to convenience and the associated costs for the educational institution.

The results obtained from the ML techniques applied to the selected sample of 14,495 undergraduate students consistently showed that the average grades in the first university period (5 weeks), the admission tests, and the average grades in high school are the three most important variables to predict undergraduate dropouts. This is in line with research reported by different authors as presented in the literature review (**Table 1**). For instance, Abu-Oda and El-Halees (2015), Hedge and Prageeth (2018), Olaya et al. (2020), and Von Hippel and Hofflinger (2020) reported that variables such as *entrance exam scores, high-school grades, University*, and *first-year academic records* are relevant for dropout prediction. Our findings are based on the study of the importance of variables obtained with the Random Forest classifier, clustering analysis, and the study of predictive power through density functions. The results from this work strongly suggest detecting students with high-risk dropout timely in the first weeks of the first academic term. Organizing additional individualized tutoring or workshops to support students with high-risk dropout characteristics during this period should be implemented as soon as possible.

Of the eight classifiers explored in this research, Random Forest (RF) provided the highest percentages of accuracy for the total sample of students, the students who dropped out, as well as for the students who were retained. The results found in this study indicate that, given the imbalance between the dropout and retention variable percentages in our student sample (8.5% and 91.5%, respectively), the best threshold value is not the one that gives the best accuracy for the whole sample to predict retentions and dropouts, but the one that gives the best precision in determining dropouts *while still* maintaining an acceptable precision in the retention and global precisions (**Figure 8**). The equilibrium point (where the three curves intersect) shows that it is possible to attain dropout precision close to 0.70 while also maintaining the retention and the global precisions at about the same value. Institutions may consider this value to better determine high-risk students and implement more focused actions to attend to this population, making the implemented resources more efficient. On the contrary, if global precision as high as 0.92 is adopted, the dropout precision would be only about 0.12 (**Table 4**), missing a great majority of high-risk students. While it is always possible to find rules that can classify any individual class with 100% precision, this comes at the expense of losing precision for another one, as shown in the Precision vs. Threshold probability graph (**Figure 8**).

It is worth noting that the database employed for this study contains dropout information only from the first academic year. Therefore, a limitation of this research is the lack of data for students dropping out in the third or higher semesters of their academic programs, although it is likely that the corresponding numbers are lower than for the first two terms because these students would be expected to be more adapted to the characteristics of their programs. More information about students after their second year is required to fully understand the long-term effect of the intervention efforts from academic, social, and economic perspectives.

Currently, we are expanding the study to include qualitative or categorical variables such as socioeconomic variables, honors and award, scholarships, social lag, parental education, etc. This second part of the research will be published elsewhere (see Gonzalez-Nucamendi et al., 2023).

The design of specific retention programs is beyond the scope of this study, and they are expected to include a comparative study of experimental (with academic intervention programs) and control groups over the years. These programs should prove the virtues of the models of improvement in tailoring retention efforts in middle and higher education over conventional predictive modeling approaches (e.g., Olaya et al., 2020).

7. Conclusion and future work

Through various Machine Learning techniques, the main variables associated with first-year undergraduate student dropouts in 14,495-student sample of the selected Case Study were identified. The most relevant numerical classification variables were: (a) the student’s academic performance in the first weeks of the first semester, (b) the average grades of the previous academic level, (c) the general entrance score in mathematics, and (d) admission test results. Other important variables included: (a) the number of class hours, (b) the age of the student, (c) the scholarship, (d) the English level, and (f) the number of subjects dropped in the first weeks of the term.

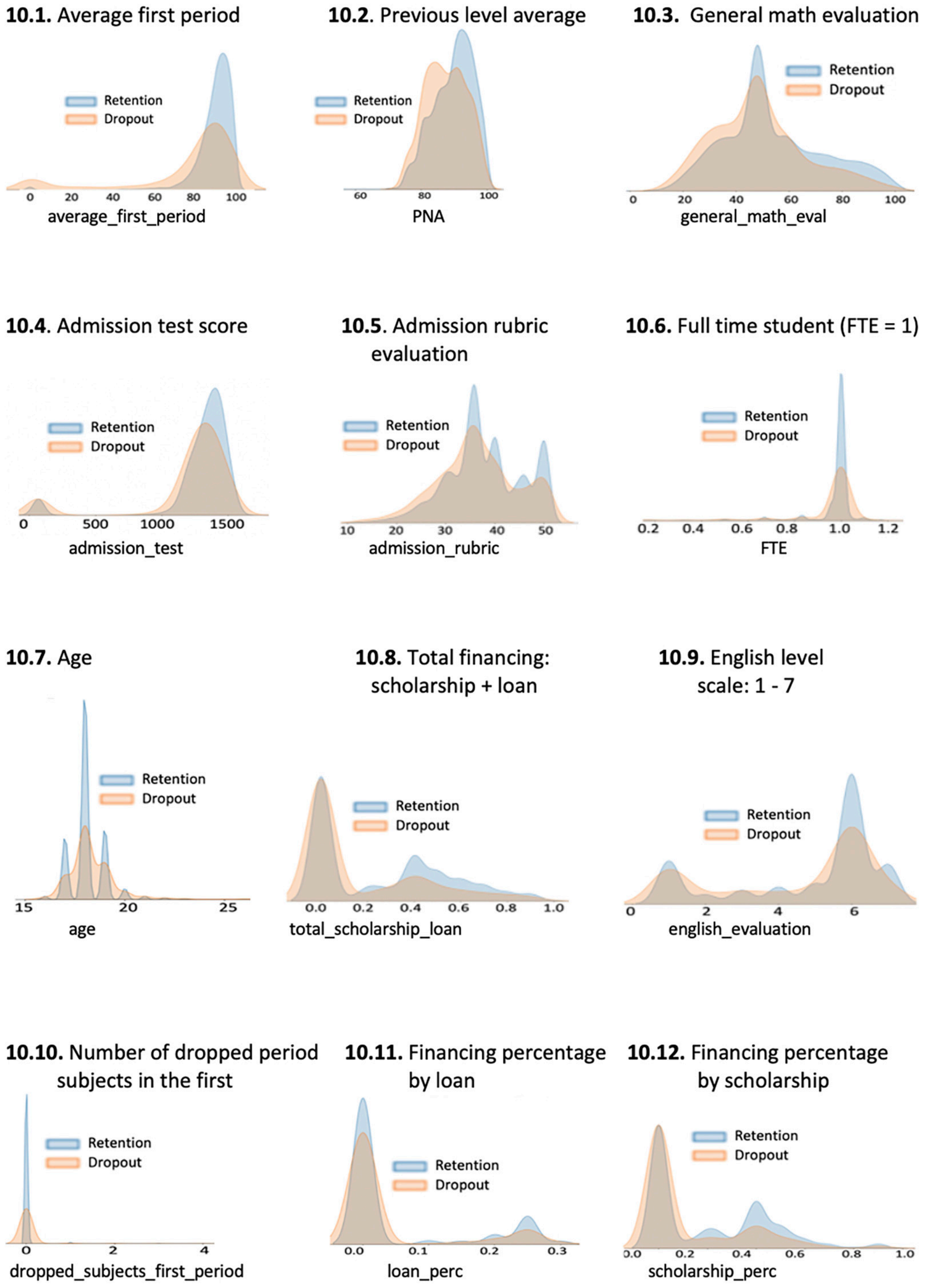

FIGURE 10
Density functions comparison for dropout and retention predictive variables.

Among the eight classifiers explored in the analysis of the Case Study data of this research, the Random Forest (RF) classifier provided the highest percentages of accuracy for the total sample of students, the students who dropped out, as well as for the students who were retained. Analyzing the predictions obtained with various classification algorithms, a negative relationship was found between the accuracies in predicting dropout and retention percentages. This led us to the use of a probability threshold different from 50% as a classification criterion to favor the smallest class and achieve a better balance in the prediction accuracy between unbalanced classes. This resulted in an improved accuracy in detecting dropouts. With this, a control is also provided that allows regulating the dropout and retention precision levels to achieve flexibility so that universities can adapt them to their objectives, resources and needs. In the database analyzed, the use of the Random Forest algorithm to implement the Threshold Probability methodology resulted in the most appropriate approach.

Consequently, the results for the Case Study of this research clearly show that the best strategy is not the one that provides the best overall prediction accuracy for the whole student sample, but the one that predicts the highest accuracy in dropout percentage while still maintaining appropriate overall and retention probabilities precision.

The design and the implementation of segmented or personalized interventions are better than random, non-focalized interventions. In this sense, academic institutions should provide appropriate programs to offer tutoring and support primarily to those students early detected as possible dropout candidates, to increase their retention probabilities.

Data availability statement

The data that support the findings of this study are available from the Institute for the Future of Education (IFE)'s Educational Innovation collection of the Tecnologico de Monterrey's Data Hub, but restrictions apply to the availability of these data, which were used under a signed Terms of Use document for the current study, and so are not publicly available. Data are however available from the IFE Data Hub upon reasonable request at https://doi.org/10.57687/FK2/PWJRSJ (accessed January 28, 2023).

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent was not required in accordance with the national legislation and the institutional requirements.

Author contributions

AG-N, JN, LN, VR-R, and RG-C contributed to the conception and design of the study and wrote the first draft of the manuscript. VR-R organized the database. AG-N performed the statistical analysis. All authors contributed to the manuscript revision, read, and approved the submitted version.

Acknowledgments

We would like to acknowledge the financial support of Writing Lab, Institute for the Future of Education, Tecnologico de Monterrey, Mexico, in the production of this work. We also would like to acknowledge Vicerrectoría de Investigación y Posgrado, the Advanced Artificial Intelligence Research Group, the Cyber Learning and Data Science Laboratory, and the Science Department of Tecnologico de Monterrey, Mexico City Campus, in the processes of this research.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1244686/full#supplementary-material

References

Abu-Oda, G. S., and El-Halees, A. M. (2015). Data mining in higher education: University student dropout case study. *Int. J. Data Mini. Knowl. Manag. Process* 5, 15–27. doi: 10.5121/ijdkp.2015.5102

Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hedge, V., et al. (2022). Student dataset from tecnologico de monterrey in Mexico to predict dropout in higher education. *Data* 7:119. doi: 10.3390/data7090119

Amare, M. Y., and Simonova, S. (2021). "Global challenges of students' dropout: A prediction model development using machine learning algorithms on higher education datasets," in *Proceeding of the 21st international scientific conference globalization and its Socio-Economic Consequences. SHS web of conferences, 129, 09001 2021*. doi: 10.1051/shsconf/202112909001

Aulck, L., Velagapudi, N., Blumenstock, J., and West, J. (2016). Predicting student dropout in higher education. *arXiv* [Preprint] arXiv:1606.06364

Balfanz, R., Herzog, L., and Mac Iver, D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. *Educ. Psychol.* 42, 223–235. doi: 10.1080/00461520701621079

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Cameron, S. V., and Heckman, J. J. (2001). The dynamics of educational attainment for black, hispanic, and white males. *J. Polit. Econ.* 109, 455–499. doi: 10.1086/321014

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, (San Francisco CA), 785–794. doi: 10.1145/2939672.2939785

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Cover, T., and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13, 21–27. doi: 10.1109/TIT.1967.1053964

Douzas, G., Bacao, F., and Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inform. Sci.* 465, 1–20. doi: 10.1016/j.ins.2018.06.056

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification*, 2nd Edn. Hoboken, NJ: Wiley.

Dynarski, M., Clarke, L., Cobb, B., Finn, J., Rumberger, R., and Smink, J. (2008). *Dropout prevention: A practice guide (NCEE 2008–4025)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Freund, Y., and Schapire, R. E. (1996). "Experiments with a new boosting algorithm," in *Proceedings of the 13th international conference machine learning*, (San Francisco, CA), 148–156.

Garg, A., Lilhore, U., Ghosh, P., et al. (2021). "Machine learning-based model for prediction of student's performance in higher education," in *Proceedings of the 8th international conference on signal processing and integrated networks, SPIN*, (Noida), 162–168.

Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., and García-Castelán, R. M. G. (2023). "Identifying the factors that affect higher education students at risk of dropping out IEEE," in *Proceedings of the frontiers in education conference (FIE)*, (College Station, TX).

Govender, C. M. (2020). Hopes, challenges and goals–voices of first-year at-risk higher education students in South Africa. *S. Afr. Rev. Sociol.* 51, 55–69. doi: 10.1080/21528586.2020.1806919

Hedge, V., and Prageeth, P. P. (2018). "Higher education student dropout prediction and analysis through educational data mining," in *Proceedings of the 2018 2nd international conference on inventive systems and control (ICISC)*, (Coimbatore), 694–699.

Heublein, U. (2014). Student drop-out from german higher education institutions. *Eur. J. Educ.* 49, 497–513. doi: 10.1111/ejed.12097

Hsu, C. W., and Yeh, C. C. (2019). Mining the student dropout in higher education. *J. Test. Eval.* 48, 4563–4575. doi: 10.1520/JTE20180021

Lamb, S., Markussen, E., Teese, R., Sandberg, N., and Polesel, J. (eds.) (2010). *School dropout and completion: International comparative studies in theory and policy*. Berlin: Springer Science & Business Media, doi: 10.1007/978-90-481-9763-7

Liu, T., Wang, C., Chang, L., and Gu, T. (2022). Predicting high-risk students using learning behavior. *Mathematics* 10:2483. doi: 10.3390/math10142483

Nagy, M., and Molontay, R. (2018). "Predicting dropout in higher education based on secondary school performance," in *Proceedings of the 2018 IEEE 22nd international conference on intelligent engineering systems (INES)*, (Las Palmas de Gran Canaria), doi: 10.1109/INES.2018.8523888

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., and Verbeke, W. (2020). Uplift modeling for preventing student dropout in higher education. *Decis. Support Syst.* 134:113320. doi: 10.1016/j.dss.2020.113320

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* [Preprint] arXiv:2010.16061

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi: 10.1007/BF00116251

Rodríguez Rojas, O. (2022). *Notas del curso, ML207 calibración de modelos en python del programa experto en machine learning de PROMIDAT generación Ada Lovelace*. Granadilla: PROMIDAT.

Romesburg, C. (2004). *Cluster analysis for researchers*. Morrisville, NC: Lulu Press.

Saravanan, T., Nagadeepa, N., and Mukunthan, B. (2022). "The effective learning approach to ICT-TPACK and prediction of the academic performance of students based on machine learning techniques," in *Communication and intelligent systems. Lecture notes in networks and systems*, Vol. 461, eds H. Sharma, V. Shrivastava, K. Kumari Bharti, and L. Wang (Singapore: Springer), doi: 10.1007/978-981-19-2130-8_7

Tec21 (2022). *Modelo Tec21*. Available online at: https://tec.mx/en/model-tec21 (accessed September 8, 2022).

Von Hippel, P. T., and Hofflinger, A. (2020). The data revolution comes to higher education: Identifying students at risk of dropout in Chile. *J. High. Educ. Policy Manag.* 43, 2–23. doi: 10.1080/1360080X.2020.1739800

Wexler, J., and Pyle, N. (2012). Dropout prevention and the model-minority stereotype: Reflections from an Asian American high school dropout. *Urban Rev.* 44, 551–570. doi: 10.1007/s11256-012-0207-4