Predictive Model for Taking Decision to Prevent University Dropout

Argelia B. Urbina-Nájera*, Luis A. Méndez-Ortega

UPAEP-Universidad (México)

Received 17 December 2020 | Accepted 2 June 2021 | Early Access 25 January 2022

unir
LA UNIVERSIDAD
EN INTERNET

Abstract

Dropout is an educational phenomenon studied for decades due to the diversity of its causes, whose effects fall on society's development. This document presents an experimental study to obtain a predictive model that allows anticipating a university dropout. The study uses 51,497 instances with 26 attributes obtained from social sciences, administrative sciences, and engineering collected from 2010 to 2019. Artificial neural networks and decision trees were implemented as classification algorithms, and also, algorithms of attribute selection and resampling methods were used to balance the main class. The results show that the best performing model was that of Random Forest with a Matthew correlation coefficient of 87.43% against 53.39% obtained by artificial neural networks and 94.34% accuracy by Random Forest. The model has allowed predicting an approximate number of possible dropouts per period, contributing to the involved instances in preventing or reducing dropout in higher education.

Keywords

Artificial Neural Network, Decision Trees, Dropouts, Educational Data Mining, Higher Education.

DOI: 10.9781/ijimai.2022.01.006

I. Introduction

Dropout in higher education is a difficult topic to explain; for this reason, [1] proposes to analyze the phenomenon from different perspectives: student, institutional, and state or national. From the student's perspective, there are expectations, goals, intellectual capacities, and socio-economic origin. In the institutional aspect, those options offered to students to include them in university life follow them up and retain them. At the state or national level, desertion must be considered the interruption of studies in any modality, and the policies that promote retention must be analyzed.

In Mexico, 38% of the people who can access higher education do not graduate, which is why the OECD (Organization for Economic Cooperation and Development) places Mexico and Turkey (with the same percentage) as the countries with a severe problem in terms of school dropouts. This percentage contrasts with Germany and Finland, where they have 4.03% and 0.45%, respectively [2]. Additionally, the effects of dropping out are reflected in labor and social inequality because the probability of finding better jobs with greater privileges is permeated by this problem [3]. According to OECD data (2018), 85% of people with higher education (25 to 64 years old) are employed, compared to 75.2% of those with only an average higher education, which shows that there are more significant job opportunities for those who advance and complete their academic studies. In Mexico's case, this proportion is 80% for those with higher education and 70.6% for those with lower education, which is below the OECD average [4].

On the other hand, according to the report on Higher Education in Mexico (2018), by the OECD, the hiring of young people between 25-34 years of age with higher education is 80.7%, which is less than 84.1% of the average of the other member countries. Thus, [3] indicates that reducing student dropout in higher education impacts positively by promoting a society better prepared to meet global challenges in which we are immersed and improve people's quality of life, get better jobs, wages, opportunities for intellectual growth, among others.

In [5], authors comment that academic desertion results from several factors such as personal situation, educational quality, facilities, socio-cultural and economic factors. This causes students to have to interrupt their studies and, therefore, affect their academic life, the institution, and society. In [1] the basis to face this problem is established and it suggests analyzing the factors in personal, institutional, and state perspectives, the first efforts to understand the desertion phenomenon was directed to explain the factors that trigger it. Recently, in [6], authors suggest applying for the latest advances in information and communication technologies (ICTs) and data science to explain this situation, not only to detect explanatory factors, but also to create predictive models that prevent it and, thus, to make decisions that reduce the mentioned indexes.

This study aims to find a model based on computational learning algorithms (decision trees and neuronal networks) that anticipates university desertion aimed at reducing the desertion rate in degree programs in Engineering, Social and Administrative Sciences.

This experimental study was based on educational data mining methodologies and computational learning algorithms such as neural networks and decision trees. The document was organized as follows: Section II describes the related works that give theoretical support to the study. Section III details the decision tree and neural network algorithms. Section IV presents the results obtained, and section V details the conclusions and future work from various approaches.

* Corresponding author.

E-mail address: argeliaberenice.urbina@upaep.mx

Please cite this article in press as:
A. B. Urbina-Nájera, L. A. Méndez-Ortega. Predictive Model for Taking Decision to Prevent University Dropout, International Journal of Interactive Multimedia and Artificial Intelligence, (2022), http://dx.doi.org/10.9781/ijimai.2022.01.006

II. Related Work

This section presents related work divided according to the method used to predict college dropout. For example, in [7][8][9][10][11], they used a statistical analysis to perform an analysis of social implications and preventive actions [8], determine actions to prevent dropout during the first semester [9] or identify factors influencing university student satisfaction, dropout, and academic performance [7][10][11].

On the other hand, various methods have also been used, such as retention theory, to identify whether the student's perspective, the institution, and the state influence university dropout [1]. In [6], heuristic and projective analysis were applied to analyze whether economic differences, vocation, attitudes, and expectations influence dropout. Also, in [5], the non-probabilistic and propositional method was used to determine whether the economic, school, and institutional core variables determine the dropout decision.

Recently, however, artificial intelligence, data mining, and computational learning methods have gained great importance in predicting college dropouts. For example, commonly used methods for predicting college dropout are: neural networks [12]-[16], K-nearest neighbors and logistic regression [14],[17], random forest [14],[17], Bayesian networks [18], decision trees [16],[19],[20], support vector machines [21], [29], statistical methods [22]-[27] and finally, deep learning in [28].

A summary of the results obtained for the best classifiers with different methods is presented below. A work [12] proposes a data mining application to make a prognosis of desertion in higher education students. The data used comprises 2007 to 2014, with 421,282 records, which the university's data warehouse provided. They analyzed data such as age, gender, location, level of studies of the tutors, year of entry to the career, and subjects failed and approved. To generate the model, they used Microsoft Azure Machine Learning's cloud service with the algorithm of Two-Class Bayes Point Much and Neural Networks. Finally, the model had an accuracy of 66%, which allowed concluding that the forecasts' results must be taken with particular caution since, although they can be improved, several factors may not be considered to assume that a way to forecast dropout was found.

In [8], authors used 5,288 student records from four generational cohorts and a decision tree model was implemented in RapidMiner Studio, with demographic variables, economic status, and some data collected at the time of entry, such as knowledge test scores. The tree used had a maximum depth of 20 and an accuracy of 87.27% to detect three factors that explain dropout: grade point average, progression period, and entrance exam score, which encourages the use of decision tree algorithms to counteract student dropout. Finally, Table I presents the works analyzed whose efforts are focused on detecting those factors that affect dropout, and few are applying computer learning methods to make a prognosis.

On the other hand, another study [13] analyzes the performance of Random Forest, Neural Networks, Support Vector Machines, and logistic regression in order to predict college student dropouts. They found that Random Forest was the best predictor by obtaining 91% of correctly classified dropouts with a sensitivity of 87%. In the study, they used a set of 80,527 records and 21 variables classified in the categories: dropout (3), demographic (4), program (7), and academic history (7).

One of the best results obtained in the analyzed articles was achieved in [14], with a data set of 61,340 records and 18 variables it was obtained a recall of 92.4% and sensitivity of 68.6% by applying logistic regression. They determined that out of the four classifiers used (Random forest, logistic regression, K-nearest neighbors, and neural networks), the classifiers logistic regression and the neural network had proven superior to all the other classifiers' highest performance metrics when the over-sampling technique was employed.

Finally, in [15] authors used a set of 2,670 records and 11 variables, applied neural networks with basis radial function and perceptron multilayer; with the first obtained an accuracy of 96.8% in training and 98.1% with test data, while with the latter achieved 96.3% with training data and 98.6% with test data, determining that the neural networks provide a model that helps to determine university dropout and with it, contributed institution administrators to make decisions before a dropout occurs.

The analysis presented in this section provides a guideline for

TABLE I. Abstract to Related Works

Year	Description	Method
Tinto (1989) [1]	Perspectives: student, institutional and state.	Retention theory
Rodríguez y Hernández (2008) [42]	Factors: work, family economy, school performance, study and orientation.	Survey and qualitative analysis
SEP (2011) [8]	Analysis of social implications and preventive actions	Statistical analysis
Moine (2013)[22]	Evaluation of methodologies and software	KDD, CRISP-DM, SEMMA y KATALYST
Report PEM 2016 [2]	Preventive actions in the first semesters	Statistical analysis
Ramírez, Espinosa y Millán 2016 [6]	Economic differences, vocation, attitudes, expectations	Heuristics and projective scope
López y Beltrán 2017 [5]	Economic, school and institutional core	Non-probabilistic and propositional
Chinkes 2017 [12]	Demographic data analysis	Machine Learning, algorithm Two-Class Bayes Point Much and artificial neural network
Carvajal, González y Sarzoza, 2017 [23]	Institutional data analysis	Descriptive, correlational and inferential statistics
Cendejas, Acuña, Cortez y Bolaños, 2017 [24]	Evaluation of methodologies and software	CRISP-DM y SEMMA
Zavala, Álvarez, Vázquez, Gonzales y Bazán, 2018 [25]	Internal student, external and bilateral factors	Correlation factors
Muñoz-Camacho, S., Gallardo,T. Muñoz-Bravo, M. y Muñoz-Bravo, C., 2018 [26]	Institutional data analysis	Logit Discrete Choice Model
Gallegos, Campos, Canales y González, 2018 [27]	Institutional data analysis	Logit probability model
Ramírez y Grandón, 2018 [19]	Analysis of demographic, economic and other data	Decision trees
Villagrá-Arnedo, et al. 2020 [29]	Performance prediction	Support Vector Machine

choosing the computational learning methods to be applied. In this case, we chose to use oversample and undersample techniques, to have a balanced data set [14],[31], as well as neural networks despite being used in [12]-[16] with a smaller number of records in some cases, our challenge is to use a more extensive data set covering an analysis period of 9 years classified in 3 areas of knowledge (engineering, social sciences, and administrative sciences) to obtain equal or better results in accuracy and sensitivity. The random forest mentioned in [14],[17] will also be applied, with the contribution in this study of using different attribute selection methods that are explained in later sections and that are not described in the analyzed articles.

III. Educational Data Mining

Educational Data Mining (EDM) is an area of computer knowledge focused on creating methods to examine the unique types of data that come from large volumes of data in educational environments to provide answers to educational questions or improve educational or administrative processes automated manner. EDM methods are drawn from various areas, including data mining, computer learning, psychometrics, statistics, information visualization, and computer modeling [31]. Currently, some algorithms have been applied in various real-world contexts to provide solutions with high precision, to mention a few, in improving a person's response time (Neurobiology), molecular regulation of alpha-viruses (Molecular Biology), geographic sales trends (Finance), monitoring of manufacturing processes (Control), classification and characterization of patients (Medicine).

These algorithms include decision trees, k-means, vector support machines, artificial neural networks, Bayesian learning, instance-based methods, and Bayesian models. In this section, decision trees and artificial neural networks are briefly described, and the metrics used to evaluate their performance.

A. Decision Trees

Decision tree algorithms are supervised learning techniques, easy to implement and very useful, composed of a single initial node and underneath other independent trees that indicate the predictive attributes [32]. The decision trees are located within a branch of automatic learning called symbolic learning, in which there are also decision rule models closely related to the trees.

Learning using decision trees is a technique that allows the analysis of sequential decisions based on the use of results and associated probabilities. An article [33] defines it as "A method of approximation of an objective function of discrete values in which a decision tree represents the objective function. Learned trees can also be represented as a set of rules ...". On the other hand, they are among the most widely used inductive learning methods in inductive inference algorithms and have been successfully applied to learning how to diagnose medical cases and assess credit risk in loan applications. It should be noted that some of the applications of this algorithm are Binary searches, expert systems, medical diagnostics, scheduling, risk analysis, among others [30].

The decision tree algorithm C5.0 (in its commercial version known as C4.5) is an extension of ID3. It can work with continuous values for the attributes, separating the possible results into two branches. The trees it generates are less leafy because each leaf does not cover a particular class but a class distribution. C5.0 forms a decision tree from the data employing recursively executed partitions, according to the depth-first strategy. Before each data partition, the algorithm considers all possible tests that can divide the data set and select the test results in the highest information gain or the highest information gain ratio. For each discrete attribute, a test with n results is considered, where n is the number of possible values the attribute can take [33].

B. Artificial Neural Network

Artificial neural networks (ANN) are computer models that try to mimic the neurons in the human brain and solve complex learning problems. They are composed of algorithms that process a set of data to find non-linear relationships. They can learn and improve their functioning [32]. The simplest type of ANN is the so-called perceptron, which takes a vector of real values as input, calculates a linear combination of these inputs, and produces a value (usually 0 or 1) according to a function. A typical ANN is formed by interconnected neurons arranged in three layers (this may vary). The data entry through the input layer passes through the hidden layer and exits through the output layer. It is worth mentioning that the hidden layer can create several layers. In other words, neurons organize in layers (monolayer and multilayer), and the output of some neurons are the inputs of other neurons, then they are forward (feedforward) if they have connections backward, then they are feedback [30].

Therefore, an artificial neural network architecture is the structure or pattern of network connections, usually grouped into structural units called layers; within a layer, neurons can be of the same type. This architecture includes three layers, the input layer that receives data or signals from the environment, the output layer that provides the network response to input stimuli, and the hidden layer that does not receive or provide information to the environment, and then they are used as internal network processing.

C. Performance of Algorithms

To estimate and compare the algorithms' performance, one of the techniques used is based on the confusion matrix (Table II).

TABLE II. Confusion Matrix

Real Values	**Prediction**	
	Positive	Negative
Positive	a	b
Negative	c	d

From Table II the metrics accuracy, precision, specificity, and recall are derived. Accuracy: percentage of true positives and negatives against all data. Precision: percentage of true positives out of the total number of classified positives. Specificity: of the true negatives, how many did you classify correctly?. Recall: of the true positives, how many did you classify correctly?.

Similarly, a study [35] recommends using the Matthew Correlation Coefficient (MCC) as a metric to evaluate the performance of binary classification models globally. It evaluates in a range of -1 to 1, where 1 represents the perfect classification, 0 a random classification, and -1 an inverse classification. Therefore, the MCC and balanced accuracy are considered in the performance evaluations. The MCC will be represented in percentage for practical purposes. The metrics to evaluate the algorithms' performance will be the same, to be able to compare and select the model with the best results.

IV. Methodology

Fig. 1 shows the method used in this study, which comprises two stages: 1) data processing and 2) classification model. The first stage bases on the knowledge discovery process (KDD), starting from describing the data set, processing, and model construction, described in points A and B of this section. In the second stage, ranking methods are applied to find the descriptive attributes. The decision trees algorithm and neural networks are applied to obtain classification models. Finally, the classification model with the best performance is obtained; this process is described in the results section.

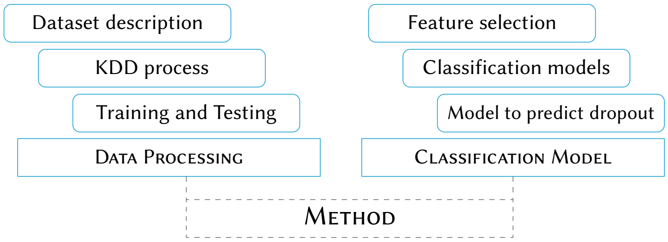


Fig. 1. The method applied to find the best predictive model of dropout.

A. Dataset Description

Table III presents the distribution of data used in this study, obtained from three deaneries distributed among 25 programs collected from fall 2010 to fall 2019 (Table III).

TABLE III. Dataset Description

Deanery	#Programs	#Records	%programs
Engineering	10	19,174	37.23
Social Science	6	11,092	21.45
Administrative Science	9	21,231	41.23
Total	25	51,497	100

TABLE IV. Attribute Description

#	Attribute	Description
1	PERIODO	Spring, summer and autumn
2	PROM_PANTE	Grade point average of the previous period
3	PROM_INI	Grade point average at the beginning of the period
4	EDAD	The age with a range between 18 to 24 years old
5	FALTAS_PANTE	Absences last period
6	ASIST_PANTE	Attendance until your last period
7	GENDER	M=Male, F= Female
8	SEMESTRE_PANTE	Semester to the period before
9	SUPPORT	Indicates if you receive any type of financial support such as discounts or agreements
10	REPRO_PANTE	Courses failed up to their last period
11	REPRO_1X	Courses failed only once
12	REPRO_2X	Courses failed twice
13	REPRO_3X	Courses failed three or more times
14	PERIODOS_INSCRITOS	All periods in which you have registered
15	PROMEDIO_AC	Current grade point average
16	CRED_PANTE	Credits taken
17	AVANCE_XCRED	Progress according to approved credits
18	ASIG_INSCRITASP	Courses registered in the period
19	FALTASP	Absences during the period
20	ESTADOCIVIL	Marital status
21	NINGRESO	This indicates that he/she is a new student
22	FORANEO	Indicates if you come from out of state
23	REPROBADAS1P	The proportion of courses failed in the first partial period
24	REPROBADAS2P	The proportion of courses failed in the second partial
25	REPROBADAS3P	The proportion of courses failed in the third partial
26	DESERTOR	Defines the deserter class in a dichotomous way (Yes or no)

Each record contains 25 attributes plus the Deserter class described in Table IV. The Deserter class labeled the data, of which Deserter(S) is 6,282, and No Deserter(N) is 45,215, so there is an unBalanced class. This class type is a problem for predictive models since it can lead to erroneous results in the prediction. That is, the class with a more significant number of samples obtains better predictive performance than the class with few samples, which is often the one of most significant interest [36]-[38].

Then, as mentioned above, the Deserter(S) class represents 12.19% of the total samples, thus exposing a class balance problem. To solve this problem, we will use the oversample and undersample techniques, because as mentioned by [30], they are easy to implement and obtain excellent results. The proposal uses an oversample for the Deserter(S) class and an undersample for the Deserter(N) class. In this way, a new set of data will be obtained with a balanced class, moving on to the training phase. The description of the attributes is shown in Table IV.

Additionally, a data set of 2,244 records concerning spring 2020 will be used without labeling to predict possible desertion in the final model.

B. KDD Process

According to [8][39][40], knowledge discovery in KDD (Knowledge Discovery in Databases) was the first model for knowledge extraction methodologically and works as a tool for decision making. In this way, the KDD process applied as follows:

The selection of attributes utilizing algorithms allows improving the input data's quality with the elimination of attributes that are not relevant [41]. In this study, we use subsection selection with CFS (Correlation Feature Selection), the filter method (Chi-square and gain information), and wrapping (Random Forest).

Pre-processing and transformation: activities such as missing data processing, noise reduction, among others, are performed. They start with data extraction from the institutional data warehouse and other sources such as Salesforce and Excel files. This information is already clean and automated by a data integration system, which employing process maps, performs the queries, validations, and transformations in an intermediate environment between the extraction and the loading of the data. All data is finally in dimension and fact tables, following the literature's standards [38]. Relational databases such as Oracle 11g, MySQL, and MSQL Server are used.

Data mining (classification model): The methods: X^2, Relief, and SOAP (Selection of Attributes by Projections) are applied since they increase accuracy, decrease overtraining since they eliminate data with better significance, and increase training speed [28]. The C5 decision tree algorithm and the artificial neural network algorithm with multilayer perceptron with 1 and 2 layer topologies with different numbers of neurons also obtain the best classification model obtained between both. For the evaluation of each algorithm's performance, the metrics of precision, accuracy, specificity, and sensitivity were used, as well as the balanced accuracy and the Matthew's Correlation Coefficient (MCC). For the validation of the models, cross-validation with base ten is employed. In the training part, the balanced subset of data applies. The proportion for the Deserter(S) class is 34.5%.

C. Training and Testing

Executions perform with different proportions in several instances; the following strategies are applied: 1) training with 70%, 80%, and 90% of the available data and 2) For the initial evaluations of the models, base ten cross-validations use.

D. Feature Selection

Two methods for feature selection are used to find the most relevant attributes. To determine the order of importance of the attributes rank methods (chi_square and gain information) were used and filter methods were used to determine a subset with the most significant attributes.

E. Classification Model

Two methods are used to build classifier models to predict attrition: 1) decision trees using the Random Forest [14],[17] method and 2) artificial neural networks [12]-[16]. Both methods are described in more detail in the results section.

F. Model to Predict Dropout

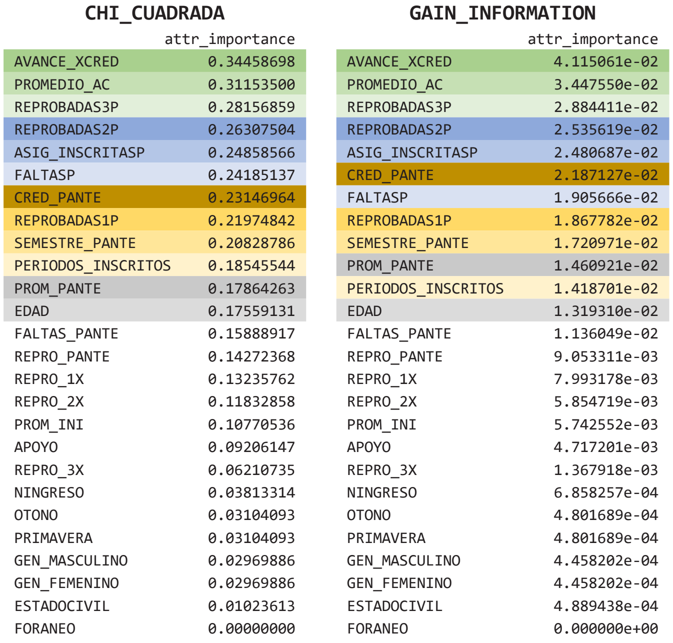
Finally, the model with the best metrics (accuracy, sensitivity, precision, among others) is chosen after performing the different experiments with test and training data. The process is described in detail in the results section.

V. Results

In this section, the results are presented in the following order: First, the descriptive attributes of the dropout phenomenon are listed; second, the predictive model obtained using decision trees is shown; third, the predictive model applying neuronal networks is presented; and finally, the model with the best performance and the prediction obtained for the spring 2020 period is described.

A. Features Selection

Two feature selection methods are applied to identify the most relevant attributes. The first method, feature selection algorithms, was used with rank methods (chi_square and gain information) to determine the most relevant attributes. In which it is observed that similar results were obtained with both methods (Fig. 2). A second method was performed with filter methods to determine a subset with the most significant attributes—also, the wrapping algorithm applies with the random forest (Table V).



CHI_CUADRADA

	attr_importance
AVANCE_XCRED	0.34458698
PROMEDIO_AC	0.31153500
REPROBADAS3P	0.28156859
REPROBADAS2P	0.26307504
ASIG_INSCRITASP	0.24858566
FALTASP	0.24185137
CRED_PANTE	0.23146964
REPROBADAS1P	0.21974842
SEMESTRE_PANTE	0.20828786
PERIODOS_INSCRITOS	0.18545544
PROM_PANTE	0.17864263
EDAD	0.17559131
FALTAS_PANTE	0.15888917
REPRO_PANTE	0.14272368
REPRO_1X	0.13235762
REPRO_2X	0.11832858
PROM_INI	0.10770536
APOYO	0.09206147
REPRO_3X	0.06210735
NINGRESO	0.03813314
OTONO	0.03104093
PRIMAVERA	0.03104093
GEN_MASCULINO	0.02969886
GEN_FEMENINO	0.02969886
ESTADOCIVIL	0.01023613
FORANEO	0.00000000

GAIN_INFORMATION

	attr_importance
AVANCE_XCRED	4.115061e-02
PROMEDIO_AC	3.447550e-02
REPROBADAS3P	2.884411e-02
REPROBADAS2P	2.535619e-02
ASIG_INSCRITASP	2.480687e-02
CRED_PANTE	2.187127e-02
FALTASP	1.905666e-02
REPROBADAS1P	1.867782e-02
SEMESTRE_PANTE	1.720971e-02
PROM_PANTE	1.460921e-02
PERIODOS_INSCRITOS	1.418701e-02
EDAD	1.319310e-02
FALTAS_PANTE	1.136049e-02
REPRO_PANTE	9.053311e-03
REPRO_1X	7.993178e-03
REPRO_2X	5.854719e-03
PROM_INI	5.742552e-03
APOYO	4.717201e-03
REPRO_3X	1.367918e-03
NINGRESO	6.858257e-04
OTONO	4.801689e-04
PRIMAVERA	4.801689e-04
GEN_MASCULINO	4.458202e-04
GEN_FEMENINO	4.458202e-04
ESTADOCIVIL	4.889438e-04
FORANEO	0.000000e+00

Fig. 2. Rank of feature selection.

Fig. 3 shows the order of attributes from least to most important with box diagrams, thus showing the dispersion of data for each. It can be noted that the variable FORANEO is not significant, and the variable REPROBADASP is significant. That means that the most significant variables are related to the failure rate, absences, semesters completed, age, average income, current average, and progress according to credits. This analysis also allowed us to detect some inconsistencies in the attributes PROM_PANTE, AVANCE_XCRED, and FALTASP:

TABLE V. The Most Relevant Attributes

#	Attributes	Wrapper-Random Forest	Consistency Based	CFS
1	ASIG_INSCRITASP	x	x	x
2	FALTASP	x	x	x
3	AVANCE_XCRED	x	x	x
4	PROMEDIO_AC	x	x	x
5	PROM_PANTE	x	x	
6	REPROBADAS3P	x	x	x
7	CRED_PANTE	x	x	
8	PROM_INI	x	x	
9	EDAD	x	x	
10	REPROBADAS2P	x	x	x
11	FALTAS_PANTE	x	x	
12	SEMESTRE_PANTE	x	x	x
13	PERIODOS_INSCRITOS	x	x	
14	REPROBADAS1P	x	x	x
15	APOYO	x	x	
16	REPRO_PANTE	x		
17	REPRO_1X	x	x	
18	OTONO	x	x	
19	PRIMAVERA	x		
20	REPRO_2X	x	x	x
21	GEN_FEMENINO	x		
22	GEN_MASCULINO	x	x	
23	NINGRESO	x	x	
24	REPRO_3X	x		
25	ESTADOCIVIL	x	x	
26	FORANEO			
	Total Attributes	25	21	9

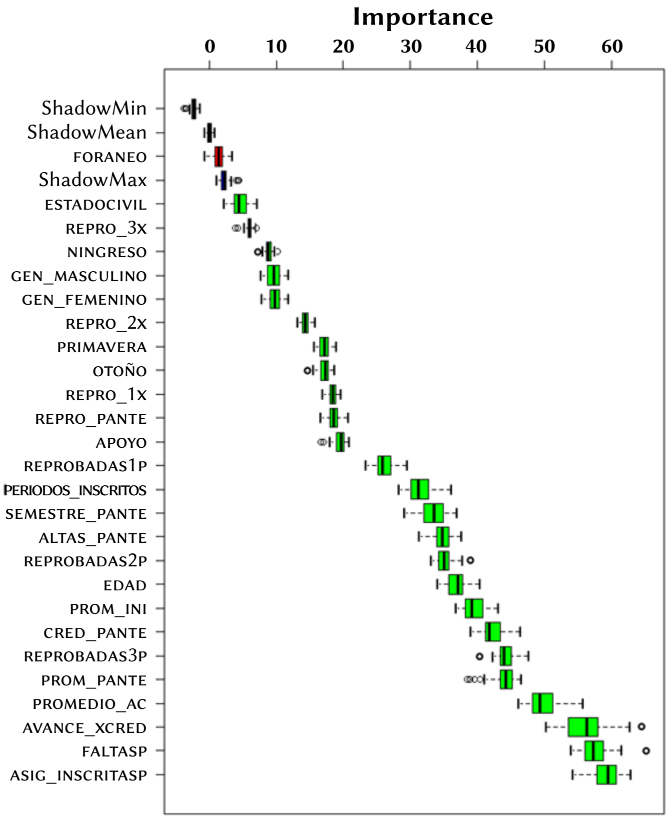


Fig. 3. Analysis with the wrapper method + random forest for the selection of attributes.

- PROM_PANTE. Most of them represent zeros values, which is not necessarily incorrect since these variables are affected by new entries and re-entries. When they are new entries, it is expected that they do not have a previous average because there is no history. The same happens with their advance by credits and faults. However, atypical data is found, where it is not typical for values greater than 0 and less than 6 to exist, since it indicates that someone who failed the previous semester and registered for the next period without credit. Thus, 90 records are discarded due to the inconsistency described.
- FALTASP. Five cases were eliminated from the data set with absences greater than 500, which could indicate an inadequate capture in the system.
- ADVANCE_XCRED. 2 cases with incorrectly calculated data from the system were discarded.

The next phase was to determine which attributes are applied for model training. For this task, decision trees with base-10 cross-validation is used as a classification algorithm. Fig. 4 shows the results with three training ratios, 70%, 80%, and 90%, and with attributes according to the methods of Correlation (CFS), Consistency, and Wrap (random-forest). It can see that the best performances were obtained with a training ratio of 90% and 80% with selection methods by Consistency and Wrap-RF. The selection of attributes by CFS causes the model to lose sensitivity with any training ratio, being the least desired option. However, it highlights that it reduces up to 9 significant attributes, which could be useful when training time must be optimized at the expense of loss of fit.

From these results (Fig. 4), it determines that the most important attributes are those obtained by the wrapper (Random Forest) due to the performance obtained. These are ASIG_INSCRITOSP, FALTASP, AVANCE_XCRED, PROMEDIO_AC, PROM_PANTE, REPROBADAS3P, CRED_PANTE, PROM_INI, AGE, REPROBADAS2P, FALTAS_PANTE, SEMESTRE_PANTE, REGISTERED_PERIODS, REPRODUCED1P, SUPPORT, PANT_PLAY, REPRODUCED1X, AUTUMN, SPRING, REPRODUCED2X, FEMALE_GEN, MALE_GEN, NINGROS, REPRODUCED3X, and STADIUM.

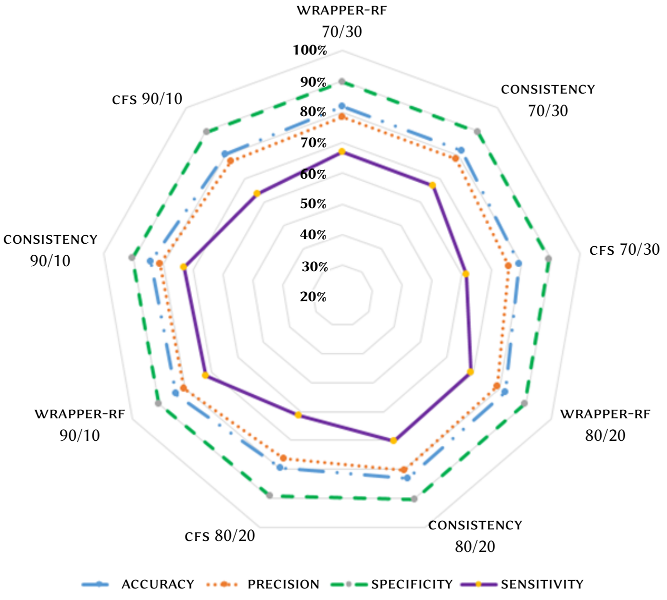


Fig. 4. Feature selection methods and their performance with decision trees.

B. Classifier Model With Random Forest Decision Trees

The decision tree algorithm C5.0 applies balanced data, and the attributes of Table IV. Fig. 4 confirms that the attributes selected by RF-Wrap give the best results, although they are very similar to those obtained by Consistency. The lowest evaluation is obtained using only the attributes by CFS. It can be seen that only the accuracy metric is lower (~88%). This metric indicates the proportion of true positives of those classified as dropouts.

The model's evaluation improvement, the balanced accuracy, and Matthew's correlation coefficient were obtained. The balanced accuracy is maintained for the attributes obtained by Consistency and RF envelope; Matthew's correlation coefficient indicates a robust positive relationship above 70%. In this way, the list of the most relevant attributes shown in Fig. 5 is obtained.

```
Attribute usage:

 100.00%  AVANCE_XCRED
  92.53%  FALTASP
  88.88%  PROMEDIO_AC
  83.00%  ASIG_INSCRITASP
  20.97%  REPRO_3X
  18.32%  REPROBADAS1P
  15.23%  REPRO_2X
  14.96%  REPROBADAS3P
   8.07%  EDAD
   7.19%  PROM_INI
   5.83%  PROM_PANTE
   4.95%  FALTAS_PANTE
   4.83%  NINGRESO
   4.02%  REPRO_1X
   2.64%  OTONO
   1.97%  PERIODOS_INSCRITOS
   1.96%  REPROBADAS2P
   1.61%  CRED_PANTE
   0.85%  ESDOCIVIL
   0.78%  APOYO
   0.48%  SEMESTRE_PANTE
   0.47%  GEN_FEMENINO
   0.35%  REPRO_PANTE
```

Fig. 5. Percentage of use of each attribute in the generation of the tree.

C. Classifier Model With Artificial Neural Network

The balanced data set obtained from the decision trees with the same attributes shown in Table V is used to train the neural network.

Two configurations of hidden layers apply, one of (2,2) and another of (12) neurons, to contrast results. The notation (2,2) refers to the number of hidden layers and neurons in the neural network architecture; the comma separates the hidden layers, the digit indicates the number of neurons per layer, so the configuration (2,2) means two hidden layers with two neurons each. At the same time, the notation (12) is one hidden layer with 12 neurons. Base 10 cross-validation was applied to evaluate the neural networks with their configurations (Fig. 6).

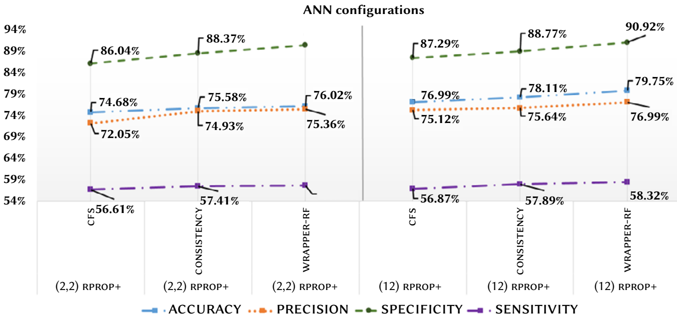


Fig. 6. Configurations for artificial neural networks with evaluation for accuracy, precision, specificity, and sensitivity.

Fig. 6 shows that the attributes obtained by CFS have the lowest values. However, as attributes are added, there is an improvement in the results. Although accuracy is high, sensitivity is low, meaning that it can detect about 57% of defectors (depending on network configuration and several attributes). On the other hand, it is analyzed the result with the Matthews correlation coefficient (MCC) and balanced accuracy.

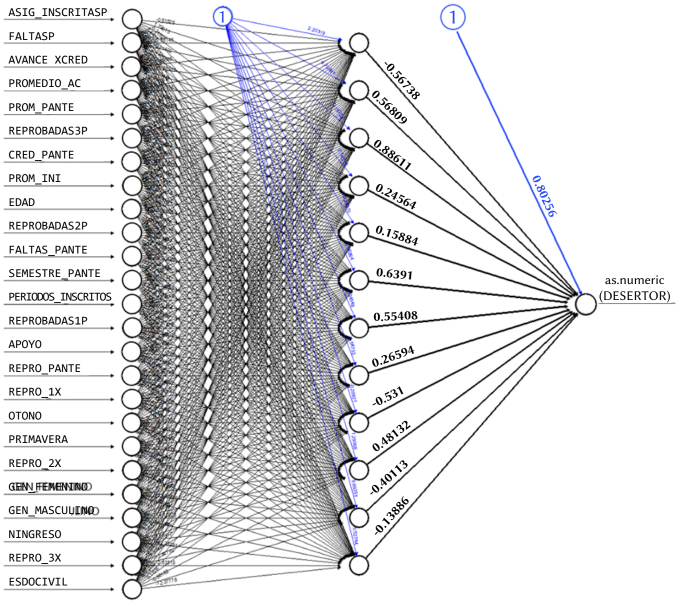


Fig. 7. A trained neural network with 12 neurons in a hidden layer.

This way, it is observed that the MMC is between 49% and 54%, while the balanced accuracy is between 71% and 75%. The Matthew coefficient indicates that the evaluations are between 49.66% and 53.69%, translating into a strong positive relationship. Besides, it can be seen that in both metrics, the minimum values are per CFS and the maximum values per Wrapper-RF. The selected model can be seen in Fig. 7 that shows only the neural network configuration with one hidden layer and 12 neurons (which resulted in the best neural network performance).

D. The Best Classifier Model

According to the classifier and configuration used, Fig. 8 shows the best results of artificial neural networks and decision trees. For the case of decision trees, the best results are given by attributes taken by RF-Wrap. For Neural Networks, the most appropriate configuration was a hidden layer with 12 neurons. The results show that decision trees are the best classification algorithm for detecting deserters.

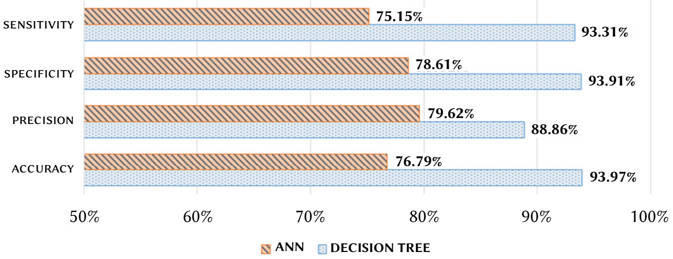


Fig. 8. Comparison between neural networks and decision trees.

Overall, when comparing the balanced accuracy and Matthew's correlation coefficient for the performance evaluation of both classification models. Similarly, Fig. 8 indicates that decision trees perform better than neural networks since decision trees achieve 94.34% balanced accuracy and a Matthew correlation coefficient of 87.43%, higher than those obtained by ANNs, 74.33%, 53.40%, respectively.

Since the objective is to detect possible defectors, an analysis was made by a period with defections from three previous periods (Fall 2018, Spring 2019, and Fall 2019) plus the current period (Spring 2020). According to the algorithm, two models were obtained per classifier and a balanced class (3x.80%) and data from autumn 2010 to spring 2018. Fig. 9 shows the actual dropouts from fall 2018 to fall 2019 and projects in spring 2020.

The prediction made by decision trees is closer to the real values from spring 2019, but it retains a uniformity in predictions between periods, while artificial neural networks are more separated from the real values than decision trees from spring 2019. This behavior is represented by the dotted line in Fig. 9. On the other hand, in the case of neural networks, training results show that, although it has better performance in the validations, Mathew's correlation coefficient is above 50%, which indicates a strong positive correlation.

Fig. 9 shows the actual results obtained at the end of the spring 2020 period. The prediction was a dropout of 436 students, and actually, 373 students dropped out in the three deaneries analyzed. The results in the Figure show the work done by the student monitoring department, which due to the Covid-19 pandemic, implemented provisional actions such as easy payment, more scholarships, agreements, personalized counseling, constant personalized monitoring, among other actions that allowed reducing the number of students predicted as possible dropouts. In this way, having a predictive model has helped to take preventive measures to reduce the dropout rate.

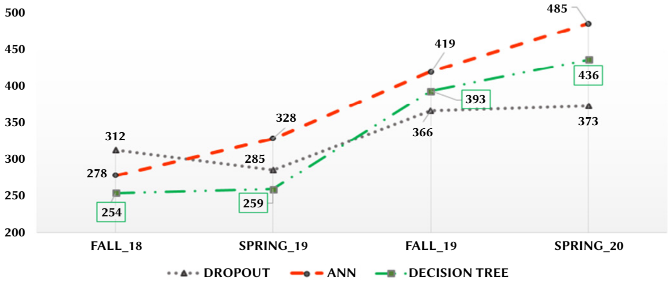


Fig. 9. Forecast of defectors and actual data obtained in the period spring 2020.

Thus, the Table VI presents a comparison of the application of neural networks. There is a difference between the number of records and variables studied in each study, determining factors in obtaining performance metrics such as accuracy and sensitivity. Although in [15], [16] the accuracy was higher than 96%, while in our study, it was 76.79%, this does not mean that neural networks are not adequate to predict dropout university in our case. Nevertheless, they are a function of the type of data, variables used, classification methods used in the neural networks, and the attribute selection methods applied before the classification, as mentioned in the explanatory features section.

TABLE VI. Comparison of Results Obtained Using Neural Networks

Reference	Dataset	# attributes	Accuracy	Sensitivity
7	43,617	Unspecified	88.30%	92.30%
32	Cohort 2010	19	82.00%	71%
35	61,340	18	87.30%	66.00%
36	2,670	11	96.80%	Unspecified
37	456	25	96.71%	Unspecified
own	51,497	25	76.79%	53.40%

The Table VII shows a comparison of the results obtained in our study with those obtained in the literature. It can be observed that the difference between accuracy and sensitivity is notorious and the number of variables and data analyzed. As in neural networks, these results' difference lies in the data set analyzed and the attributes. In summary, the closer the percentage of metrics is to 100%, the better the classifier.

TABLE VII. Comparison of Results Obtained Using Random Forest

Reference	Dataset	# attributes	Accuracy	Sensitivity
33	32,538	784	62.24%	69.40%
35	61,340	18	90.70%	68.40%
own	51,497	25	94.34%	87.43%

VI. Conclusions

In the first analysis of attributes, some that do not seem so relevant can be discarded, such as failed subjects, period (autumn or spring), and gender. However, comprehensive analysis with attribute selection by envelope - random forest shows that the models' maximum performance is obtained using most of the attributes. On the other hand, the class balance allowed us to improve the performance metrics of both algorithms. Mathew's correlation coefficient and balanced accuracy provided a better evaluation of the models, allowing the results to be unaffected by variations in accuracy, precision, specificity, and sensitivity metrics.

Decision trees obtained the best Matthew correlation coefficient of 87.43% and balanced accuracy of 94.34%. On the other hand, the tests performed indicate that increasing the number of neurons in the hidden layer of ANN could improve performance. However, it requires more processing power since server training can be more than an hour-long, and it is the main reason for not having performed more tests.

Since the best model has been the one obtained by decision trees, it has been implemented in the Enterprise Resource Planning (ERP) institutional system, which helps the student to receive the accompaniment he requires so as not to interrupt his studies; this effort translates into monitoring about 450 students per period belonging to the deaneries of social sciences, administration, and engineering. In this way, the student monitoring area improves its administrative tasks using the model because it obtains a list in a matter of seconds and avoids consolidating a report of several Excel documents, even from different areas (admissions and school control).

Dropout is a challenge for any educational institution, and with the help of algorithms and technological platforms, it is possible to collaborate in decision making to avoid dropouts or abandonment, manage better tutoring and student support. It allows better management of academic and economic resources of the institution at the institutional level, optimizing its processes and reducing response times.

To continue with the research, we have considered using additional attributes such as payment history, debts, campus access, and similar, as well as implementing stratified sampling by deanery for class balancing in the model's training and dividing data into new entries and re-entry. Also, we intend to use other classification algorithms such as near neighbors, vector support machines, logistic regression, and use a combination of classifiers to generate a more robust solution. Also, we plan to use cloud services such as Machine Learning in AWS, Azure Machine Learning, BigML, IBM Watson, TensorFlow, or some other computer learning solution to improve processing times and, finally, deploy models to classify online mode.

References

[1] V. Tinto. "Definir la deserción: una cuestión de perspectiva". Revista de Educación Superior, vol. 71, no. 18, pp. 1-9, 1989. Available: https://bit.ly/3mptENw.

[2] Organization for Economic Cooperation and Development. "Panorama de la Educación, indicadores de la OECD 2019", Ministerio de Educación y Formación Profesional, 2019. Available: https://bit.ly/3afjDjq.

[3] C. Alonso, E. Blanco, T. Fernández, and P. Solís. "Caminos desiguales: trayectorias educativas y laborales de los jóvenes en la ciudad de México". INEE-El Colegio de México. Primera edición, 2014. Available: https://bit.ly/3oVxVtt.

[4] Organization for Economic Cooperation and Development. "Higher Education in Mexico: Labour Market Relevance and Outcomes", OECD Publishing, Paris, 2019, https://doi.org/10.1787/9789264309432-en.

[5] L. López Villafaña, and A. Beltrán Solache. "La deserción en estudiantes de educación superior: tres percepciones en estudio, alumnos, docentes y padres de familia". Pistas Educativas, no. 126, 2017. Available: https://bit.ly/37qBvWJ.

[6] E. Ramírez, D. Espinosa, and E. Millán. "Estrategia para afrontar la deserción universitaria desde las tecnologías de la información y las comunicaciones". Revista Científica, vol. 24, pp. 52-62, 2016, doi: 10.14483/udistrital.jour.RC.2016.24.a5.

[7] J. R. Casanova, A. Cervero, J. C. Núñez, L. S. Almeida, and A. Bernardo, "Factors that determine the persistence and dropout of university students", Psicothema, vol. 30, no. 4, 2018, pp. 408-414, doi: 10.7334/psicothema2018.155.

[8] Secretaría de Seguridad Pública. "Deserción escolar y conductas de riesgo en adolescentes". Subsecretaría de prevención y participación ciudadana, Dirección general de prevención del delito y participación ciudadana. gobierno federal SSP. Dirección General de Prevención del Delito y Participación Ciudadana. Gobierno Federal SSP, 2011. Available: https://bit.ly/34kOwz1.

[9] Secretaría de Educación Pública. "Principales cifras del sistema educativo nacional 2015-2016, cifras preliminares". Biblioteca de publicaciones oficiales del gobierno de México, 2016. Available: https://bit.ly/3gS8juK.

[10] I. W. Li, D. R. Carroll, "Factors influencing university student satisfaction, dropout, and academic performance: an Australian higher education equity perspective," National Centre for Student Equity in Higher Education, pp. 3-59, 2017. https://doi.o10.1080/1360080x.2019.1649993.

[11] P. Perchinunno, M. Bilancia, D. Vitale, "A Statistical Analysis of Factors Affecting Higher Education Dropouts", Social Indicators Research, 2019, https://doi.org/10.1007/s11205-019-02249-y.

[12] E. Chinkes, "Pronósticos y data mining para la toma de decisiones, pronóstico sobre la deserción de alumnos de una facultad". Cuadernos del CIMBAGE, no. 20, pp. 107-132, 2017. Available: https://bit.ly/2MEyGK0.

[13] M. Solis, T. Moreira, R. Gonzalez, T. Fernandez and M. Hernandez, "Perspectives to Predict Dropout in University Students with Machine Learning," 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI), San Carlos, Costa Rica, 2018, pp. 1-6, doi: 10.1109/IWOBI.2018.8464191.

[14] N. Mduma, K. Kalegele, D. Machuve, "Machine learning approach for reducing students dropout rates. International journal of advanced computer research (IJACR)," Vol 9. No. 42, 2019. https://doi.org/10.19101/ijacr.2018.839045.

[15] M. Alban, D. Mauricio, "Neural networks to predict dropout at the universities. International journal of machine learning and computing," Vol. 9. No 2. Pp. 149-153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.

[16] N. Lázaro, Z. Callejas, D. Griol, "Predicting computer engineering students' dropout in Cuban higher education with pre-enrolment and early performance data," Journal of technology and science education. vol. 10 no. 2, 2020, doi:10.3926/jotse.922.

[17] L. Auluck, N. Velagapudi, J. Blumenstock, J. West, "Predicting Student Dropout in Higher Education," 2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, NY, USA, 2017, pp. 16-20.

[18] C. Lacave, A. Molina, J. Cruz-Lemus, "Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks", Behaviour & Information Technology, vol. 37, 2018, https://doi.org/10.1080/0144929X.2018.1485053.

[19] P. Ramírez, and E. Grandón, "Predicción de la deserción académica en una universidad pública chilena a través de la clasificación basada en árboles de decisión con parámetros optimizados," Formación Universitaria, vol. 11, no. 3, pp. 3-10, 2018

[20] A. B. Urbina-Nájera, J. C. Camino-Hampshire, R. Cruz-Barbosa, "Deserción escolar universitaria: Patrones para prevenirla usando

minería de datos educativa," RELIEVE, vol. 26, no. 1, 2020, http://doi.org/10.7203/relieve.26.1.16061.

[21] D. Sun, Y. Mao, J. Du, P. Xu, Q. Zheng, H. Sun, "Deep learning for dropout prediction in MOOCs," 2019 Eighth International Conference on Educational Innovation through Technology (EITT). Biloxi, MS. USA, 2019, doi: 10.1109/EITT.2019.00025.

[22] J. M. Moine, "Metodologías para el descubrimiento de conocimiento en bases de datos: un estudio comparativo," Universidad Nacional de la Plata. Facultad de Informática. Tesis presentada en la Facultad de Informática de la Universidad Nacional de la Plata. 2013. Available: https://bit.ly/3wbmfH0.

[23] C. Carvajal, J. González, S. Sarzoza, "Variables sociodemográficas y académicas explicativas de la deserción de estudiantes en la facultad de ciencias naturales de la universidad de playa ancha (chile)," Formación Universitaria, vol. 11, no. 2, 2017, pp. 3-12. https://doi.org/10.4067/s0718-50062018000200003.

[24] V. Cendejas, L. Acuña, M. Cortez, J. Bolaños, "El uso de modelo y metodologías de minería de datos para la inteligencia de negocios," Revista de sistemas computacionales y TIC'S. vol.3, no. 8, 2017, pp. 54-63. Available: https://bit.ly/2ThbtQV.

[25] M. Zavala, M. Álvarez, M. Vázquez, M., I. González and A. Bazán, "Factores internos, externos y bilaterales asociados con la deserción en estudiantes universitarios," Interacciones Revista de Avances en Psicología, vol. 4. no.1, 2018, pp. 59-69. https://doi.org/10.24016/2018.v4n1.103.

[26] S. Muñoz-Camacho, T. Gallardo, M. Muñoz-Bravo, C. Muñoz-Bravo, "Probabilidad de deserción estudiantil en cursos de matemáticas básicas en programas profesionales de la Universidad de los Andes Venezuela," Formación Universitaria, vol. 11, no. 4, 2018, pp. 33-42. https://doi.org/10.4067/s0718-50062018000400033.

[27] J. Gallegos, N. Campos, K. Canales, K., E. González, "Factores determinantes en la deserción universitaria. caso facultad de ciencias económicas y administrativas de la universidad católica de la santísima concepción," Formación Universitaria, vol. 11, no. 3, 2018, pp. 11-18. https://doi.org/10.4067/s0718-50062018000300011.

[28] R. Ruiz., J. Aguilar, and J. Riquelme, "Evaluación de Rankings de Atributos para Clasificación," Departamento de Lenguajes y Sistemas Informáticos. Universidad de Sevilla, Sevilla, España, 2002. Available: https://bit.ly/38bscde.

[29] C. L. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, R. Satorre-Cuerda, P. Compañ-Rosique, and R. Molina-Carmona, "Time-Dependent Performance Prediction System for Early Insight in Learning Trends," International Journal of Interactive Multimedia and Artificial Intelligence. vol. 6, no. 2, 2020. Doi: 10.9781/ijimai.2020.05.006.

[30] I. Witten, E. Frank, M. Hall, and C. Pal, "Data mining: Practical machine learning tools and techniques," Morgan Kaufmann Publishers, Burlington.

[31] C. Romero, and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135-146, 2007. https://doi.org/10.1016/j.eswa.2006.04.005.

[32] C. Menes, G. Arcos, P. Moreno, and C. Gallegos C. "Desempeño de algoritmos de minería en indicadores académicos: Árbol de decisión y Regresión Logística," Revista Cubana de Ciencias Informáticas, vol. 9, no. 4, 2015.

[33] T. Mitchell, "Decision Tree Learning," Washington State University, 2000. Available: https://bit.ly/2N1AI32.

[34] N. Sánchez, "Máquinas de soporte vectorial y redes neuronales artificiales en la predicción del movimiento USD/COP spot intradiario," Universidad Externado de Colombia, vol. 9, pp. 113-172, 2016, doi: http://dx.doi.org/10.18601/17941113.n9.04.

[35] D. Chicco, and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 6, pp. 1-13, 2020. https://doi.org/10.1186/s12864-019-6413-7.

[36] R. Malhotra, and S. Kamal, "An empirical study to investigate oversampling methods for improving software defect prediction using imbalanced data," Neurocomputing, vol.343, no. 28, 2019, pp. 120-140. https://doi.org/10.1016/j.neucom.2018.04.090.

[37] X. Zhang; Z. Shi; X. Liu; X. Li, "A Hybrid Feature Selection Algorithm for Classification Unbalanced Data Processing," IEEE International Conference on Smart Internet of Things (SmartIoT), 17-19 August 2018, doi: 10.1109/SmartIoT.2018.00055.

[38] I. Bolodurina, A. Shukhman, D. Parfenov, A. Zhigalov, and L. Zabrodina, "Investigation of the problem of classifying unbalanced datasets in identifying distributed denial of service attacks," Journal of Physics: Conference Series, vol. 1679, 2020, doi:10.1088/1742-6596/1679/4/042020.

[39] M. Rogalewicz, and R. Sika, "Methodologies of knowledge discovery from data and data mining methods in mechanical engineering," Management and Production Engineering Review, vol. 7, no. 4, pp. 97-108, 2016. https://doi.org/10.1515/mper-2016-0040.

[40] C.R.M. Rosa, M.T.A. Steiner, and P.J. Steiner Neto, "Knowledge Discovery in Data Bases: a Case Study in a Private Institution of Higher Education," IEEE Latin America Transactions, vol. 16, no. 7, pp. 2027-2032, 2018, doi: 10.1109/TLA.2018.8447372.

[41] J. S. Aguilar-Ruiz, and J. Díaz-Díaz, "Selección de atributos relevantes basada en bootstrapping," Departamento de lenguajes y sistemas informáticos. Universidad de Sevilla. ETS ingeniería informática. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, 2005, pp. 21-30.

[42] J. Rodríguez, J. Hernández, J., "La deserción escolar universitaria en México, la experiencia de la universidad autónoma metropolitana campus Iztapalapa," Revista Actualidades Investigativas en Educación, vol. 8, no. 1, 2011, pp. 1-31, doi: 10.15517/AIE.V8I1.9308.

Argelia Berenice Urbina Nájera

She is a full-time research professor at the Popular Autonomous University of the State of Puebla (UPAEP), México. She obtained a PhD from the UPAEP in 2015. She is a member of the National System of Researchers (SNI) of México since 2017. Her research focuses on Educational Data Mining, Learning Analytics, Machine Learning applied to health, business and education; as well as, other knowledge areas related to Education and Business Intelligence.

Luis Andrés Méndez Ortega

Master in Data Science and Business Intelligence from the Popular Autonomous University of the State of Puebla (UPAEP). Degree in Computer Engineering from the Autonomous University of Tlaxcala (UAT). With more than 9 years in the development and implementation of administrative software for different sectors such as automotive, motor transport, medical, commerce and government projects. He has also participated as an analyst in the area of systems for data migration, and user requirements management in the education sector. He is currently a business intelligence consultant in the CRM & BI area at UPAEP and responsible for BI architecture. Main interests in data management, ETL processes, computer learning and software development.