Prediction of student attrition risk using machine learning

Mauricio Barramuño and Claudia Meza-Narváez
Carrera Kinesiología, Sede Temuco, Universidad Autónoma de Chile, Temuco, Chile, and
Germán Gálvez-García
Departamento de Psicología, Universidad de La Frontera, Temuco, Chile and Laboratoired'Étude des Mécanismes Cognitifs, Département de Psychologie Cognitive, Sciences Cognitives et Neuropsychologie, Institut de Psychologie, Université Lyon 2, Lyon, France

Received 24 February 2021
Revised 1 May 2021
Accepted 3 May 2021

Abstract

Purpose – The prediction of student attrition is critical to facilitate retention mechanisms. This study aims to focus on implementing a method to predict student attrition in the upper years of a physiotherapy program.
Design/methodology/approach – Machine learning is a computer tool that can recognize patterns and generate predictive models. Using a quantitative research methodology, a database of 336 university students in their upper-year courses was accessed. The participant's data were collected from the Financial Academic Management and Administration System and a platform of Universidad Autónoma de Chile. Five quantitative and 11 qualitative variables were chosen, associated with university student attrition. With this database, 23 classifiers were tested based on supervised machine learning.
Findings – About 23.58% of males and 17.39% of females were among the attrition student group. The mean accuracy of the classifiers increased based on the number of variables used for the training. The best accuracy level was obtained using the "Subspace KNN" algorithm (86.3%). The classifier "RUSboosted trees" yielded the lowest number of false negatives and the higher sensitivity of the algorithms used (78%) as well as a specificity of 86%.
Practical implications – This predictive method identifies attrition students in the university program and could be used to improve student retention in higher grades.
Originality/value – The study has developed a novel predictive model of student attrition from upper-year courses, useful for unbalanced databases with a lower number of attrition students.

Keywords Student attrition, Supervised machine learning, Data classification, University student
Paper type Research paper

Introduction

Student attrition (SA) can be defined as the premature abandonment of a study program before acquiring the title or degree (Braxton, 2019; Himmel, 2002). In higher education, there is a high rate globally. For example, in the USA, it is an issue that affects around 50% of students (Stinebrickner and Stinebrickner, 2007). We encountered similar numbers in Latin America (Bruneforth *et al.*, 2004; Patiño and Cardona, 2012). In Chile, the attrition rate for university students is 21% in the first year (SIES, 2020) and increases to 31% in the second year (CNED, 2020).

SA is a complex problem comprised of various elements and perspectives (Tinto, 1993). These include individual socioeconomic, institutional and academic factors (Patiño and Cardona, 2012). Several authors propose that academic performance is crucial to predicting this phenomenon (Cuji *et al.*, 2017; Méndez, 2016; Patiño and Cardona, 2012). This could be

Financing: This study was financed by the Teaching Innovation and Development Fund from the Vice-Rector for Academic Affairs and the Vice-Rector of Research and Graduate Studies at the Universidad Autónoma de Chile.
Conflict of interest: The authors declare no conflict of interest.

Journal of Applied Research in Higher Education

2050-7003
DOI 10.1108/JARHE-02-2021-0073

determined by other modulators such as age, previous learning experiences, schools where the person studied, gender, socioeconomic situation, family environment, established interpersonal relations, makeup of groups and/or self-perception of qualities (Villamizar and Romero, 2011).

All of this establishes theoretical and explanatory parameters for this phenomenon. Although the different reasons for SA are known, however, the tools used are not sufficiently accurate to identify *a priori* that a student is at risk (Amaya *et al.*, 2014).

Because of the efforts of higher education institutions to carry out the digitization of their students' data, access to them has been facilitated, generating new opportunities for their analysis (Sandoval *et al.*, 2018). In this sense, data mining becomes important and emerges as an interesting tool to answer complex questions in education such as learning, the prediction of academic performance and SA (Mduma *et al.*, 2019).

There are currently statistical computer methods that take advantage of the information on large databases to generate predictive classifiers (Agarwal *et al.*, 2012; Amaya *et al.*, 2014; Cuji *et al.*, 2017; Dutt *et al.*, 2015). The studies conducted have used various statistical and predictive tools. Kuna *et al.* (2010) used machine learning based on data mining, specifically decision trees, with databases of first- and second-year students doing a bachelor in engineering. They obtained that the variables "who finances the studies" and "number of years completed between leaving high school and entering university" are important for the classification; however, they do not specify the accuracy achieved. Méndez (2016), using first-year undergraduate student records, tested white-box classification algorithms with induction rules and decision trees, obtaining an accuracy of up to 88.35%. Miranda and Guzmán (2017) describe success constraints in the classification of 76% with a Bayesian network, 75% with a decision tree and 83% with a neural network. Costa *et al.* (2017), using a database of 262 students and a support vector machine, obtained up to 92% accuracy in the early prediction of students' academic failure in introductory programming courses. For their part, Murakami *et al.* (2019) found up to 95% accuracy using logistic regression. This test is more accurate when using data from both current and graduate students than when only examining data from graduate students. In the same way, Adejo and Connolly (2018), using a database of 141 samples, predicted the student academic performance and related it to retention, achieved an accuracy of 79% through a hybrid model. Most of the studies in this area focus on SA in the first or second year after entry, which is the population at greatest risk of attrition (Canales and De los Ríos, 2007; Himmel, 2002), but there is a paucity of literature that addresses this subject in upper-year courses (González and Uribe, 2018). Therefore, it is not clear how these classification tools work when only records of upper-year students are used, where there is a greater imbalance in the proportion of attrition and non-attrition.

Several research studies focused on using algorithms to predict university students' attrition, specifically with samples that included students in their upper-year courses. Bedregal-Alpaca *et al.* (2020) conducted a research with engineering students using decision trees and neural networks. They found ranges between 65 and 83% of accuracy. In this avenue, Viloria *et al.* (2019) found 72% of accuracy for decision trees, 73% for neutral networks and 76% for Bayesian Network. Cuji *et al.* (2017) found an accuracy of up to 94% with a decision tree for students of computer science teaching. It should be noted that all the studies above were carried out based on university programs associated with engineering or computer science and not in the context of health students. In this vein, a research with nursing students was carried out with a decision tree (Moseley and Mead, 2008). The authors revealed a sensitivity of 84%, a specificity of 70% and an accuracy of 94%.

The purpose of this study was to establish a statistical classifier that could predict which upper-year students pose a risk of early attrition. The main difference concerning previous studies is emphasizing the construction of a training matrix composed only by upper-year students, besides not belonging to the traditionally studied areas such as engineering, mathematics or computer science. To do this (due to the accessibility to the sample), we chose

to study the physiotherapy program in a Chilean university. At a global level, the tool proposed here will establish a classification model that is novel in the literature on this topic. At a local level, it will help us establish a support system and foster student retention in the program analyzed in this study.

Materials and methods

Participants

An encoded and anonymized record of 336 students in upper-year courses in the physiotherapy program at the Universidad Autónoma de Chile was used, which represents the total number of students enrolled in the entry cohorts between 2012 and 2017. Each of the records was labeled according to attrition (Table 1). The data were collected from the Financial Academic Management and Administration System (SAGAF in Spanish) and the platform "Understanding to Include University Student Diversity" (CIDEU, by its acronym in Spanish). The use of the databases was authorized by the Vice-Dean of the Faculty of Health Sciences. The project was approved by the Scientific Ethics Committee (n°120-18) of the Universidad Autónoma de Chile, and developed according to the Declaration of Helsinki (Asociación Médica Mundial, 2000).

Selection of variables and data transformation

Based on previous research, sociodemographic (Aulck *et al.*, 2016; Chen *et al.*, 2018; Cuji *et al.*, 2017; Gil *et al.*, 2020; Ortiz-Lozano *et al.*, 2018; Siri, 2015), academic (Aulck *et al.*, 2016; Bedregal-Alpaca *et al.*, 2020; Chen *et al.*, 2018; Cuji *et al.*, 2017; Gil *et al.*, 2020; Ortiz-Lozano *et al.*, 2018; Siri, 2015; Viloria *et al.*, 2019; Wan Yaacob *et al.*, 2020), financial aid (Strecht *et al.*, 2015) and parents' educational level (Gallegos *et al.*, 2018) variables were chosen. Five quantitative and 11 qualitative variables were chosen and associated with university students' attrition, presented in Table 2. They were encoded according to their nominal, ordinal or quantitative nature and included in a unique numerical matrix. Following the stages of "Machine learning" (Fayyad *et al.*, 1996; Harrington, 2012), a ranking of these variables was generated according to the entropy of the data through the function "InfoGainAttributeEval" of the WEKA machine learning software (Frank *et al.*, 2016).

Algorithms

A brief description of the algorithms used more frequently to classify attrition students is presented as follows: (1) Regression: this is a way of fitting data to a model. A model can be a curve in multiple dimensions. The regression process fits the data to the curve, producing a model that can be used to predict (Paluszek and Thomas, 2019). (2) Neural networks: artificial neurons learns through repeated trials how to organize themselves better. They are composed of nodes (neurons) connected to the next set of nodes by a series of weighted trajectories. The prediction error is evaluated, and the weights are modified to improve the prediction in a

$n = 336$	Number of students	Dropouts	Percentage of attrition
Cohort 2012	67	18	26.8%
Cohort 2013	35	5	14.3%
Cohort 2014	38	10	26.3%
Cohort 2015	65	11	16.9%
Cohort 2016	66	10	15.2%
Cohort 2017	66	10	15.2%
Total	336	64	19.0%

Note(s): All students within each cohort are from the same semester

Table 1. Number and percentage of dropouts according to student cohort

Variable	Operational definition	Ranking
Average grades	Grade between 1.0 (lowest) and 7.0 (highest)	1
Type of financing	a. Own. b. Scholarship. c. External loan. d. Internal loan from the university. e. Free. f. Others	2
Number of subjects with double fail	Number of subjects with double fail	3
Number of subjects failed for the first time	Number of subjects failed for the first time	4
Feels family support	Yes–No	5
First in the family to enter higher education	Yes–No	6
Gender	Male–Female	7
Age	Age in years completed	8
Who they live with	a. Alone. b. With family. c. With partner/friends	9
School origin	a. Municipal. b. Subsidized. c. Private	10
Years to enter university	Number of years completed between high school and entry to university	11
Works	Yes–No	12
Mother's education level	a. Elementary complete. b. Secondary complete. c. Higher complete. d. Graduate studies complete	13
Origin	Urban–Rural	14
Father's education level	a. Elementary complete. b. Secondary complete. c. Higher complete. d. Graduate studies complete	15
Financial status	In arrears–No arrears	16

Table 2.
Ranking of variables according to entropy

Note(s): Entropy ranking obtained through the function "InfoGainAttributeEval" of the machine learning software WEKA

cyclical process (Paluszek and Thomas, 2019). (3) Vector support machine: it is a model that tries to separate the different classes by means of a space or hyperplane as wide as possible, classifying according to their proximity (Harrington, 2012). (4) Decision trees: they are algorithms for classifying using successive partitions, where a first decision must be made about the dataset to dictate which function is used to divide the data (Harrington, 2012). On the other hand, these decision trees can be enhanced by random under-sampling boosting (RUSBoost). Rus stands for random under-sampling. The algorithm takes N, namely, the number of members in the class with the fewest members in the training data, as the basic unit for sampling. Classes with more members are under sampled by taking only N observations of every class. In other words, if there are K classes, then, for each weak learner in the ensemble, RUSBoost takes a subset of the data with N observations from each of the K classes. The boosting procedure follows the procedure in adaptive boosting for multiclass classification for reweighting and constructing the ensemble (Mathworks, 2021; Seiffert *et al.*, 2010).

Training the classifier

Using the machine learning application of the Matlab® software (version R2012b, USA), 23 possibilities of statistical classifiers were tested with different hyperparameters, all based on supervised machine learning. For each, a training matrix was created with 100, 50 and 25% of the variables with greatest entropy. In addition, an analysis was included using only the first variable from this ranking.

Design and statistical analysis

This work was developed with a descriptive approach to determine the most sensitive model to predict attrition. In addition, through a quantitative approach, the following hypothesis is contrasted: There are significant differences when constructing the classification model with

the first variable, 25, 50 and 100% of the variables of the entropy ranking. For the statistical analysis, the effectiveness of each classifier was determined by calculating their overall accuracy using ten-fold cross-validation. Training and evaluation subgroups were generated according to the attrition label known *a priori* from the data matrix. Then, the arithmetic mean of the results was taken from each iteration to obtain the overall accuracy (Kohavi, 1995). The contingency table was constructed with true positives (TP), false positives (FP), true negatives (TN), false negatives (FN) and the accuracy, sensitivity, specificity and *F*-score were determined.

Overall accuracy (ACC) measures the relationship between correct predictions and the total number of cases. It was calculated according to the following equation:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

The sensitivity (SEN) corresponds to the students who were attrition and were correctly classified. It was calculated using the following equation:

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

The specificity (SPC) of a test is the probability that a non-attrition student will have a negative test result. It was calculated using the following equation:

$$\mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

*F*1 is a harmonic mean that combines the accuracy and sensitivity values, allowing the comparison and choice of the best predictive model. It was calculated using the following equation:

$$F1 = 2 \cdot \frac{\mathrm{ACC} \cdot \mathrm{SEN}}{\mathrm{ACC} + \mathrm{SEN}}$$

ROC area, this index can be interpreted as the probability that a classifier scores a positive instance chosen at random higher than a negative one.

For inferential statistics, the SPSS (Statistical Package for the Social Sciences) version 23.0 was used. The differences in accuracy were evaluated through the Kruskal–Wallis test. One-factor ANOVA test to evaluate the differences in ROC area and *f*-measure. Games–Howell's *post hoc* multiple comparisons test or the HSD Tukey test were used based on the homoscedasticity results of the Levene test. For all cases, a significance value of $p < 0.050$ was established.

Results

The average age of the group for both the attrition and non-attrition groups was 22 years. About 23.58% of men and 17.39% of women corresponded to attrition students. About 60.93% of the attrition group came from subsidized education, and 43.75% came from rural families.

The mean accuracy of the classifiers studied increased in terms of the number of variables used for the training; 78.61% using the first variable of the entropy ranking, 83.29% with 25% of the variables, 83.67% with 50% of the variables and 83.82% with 100% of the variables. Figure 1 presents the box plots and the hypothesis test. For accuracy, ROC value and *F*-measure, the box plot shows lower values when using only the first variable in the entropy ranking. In the hypothesis test, significant differences were found when constructing

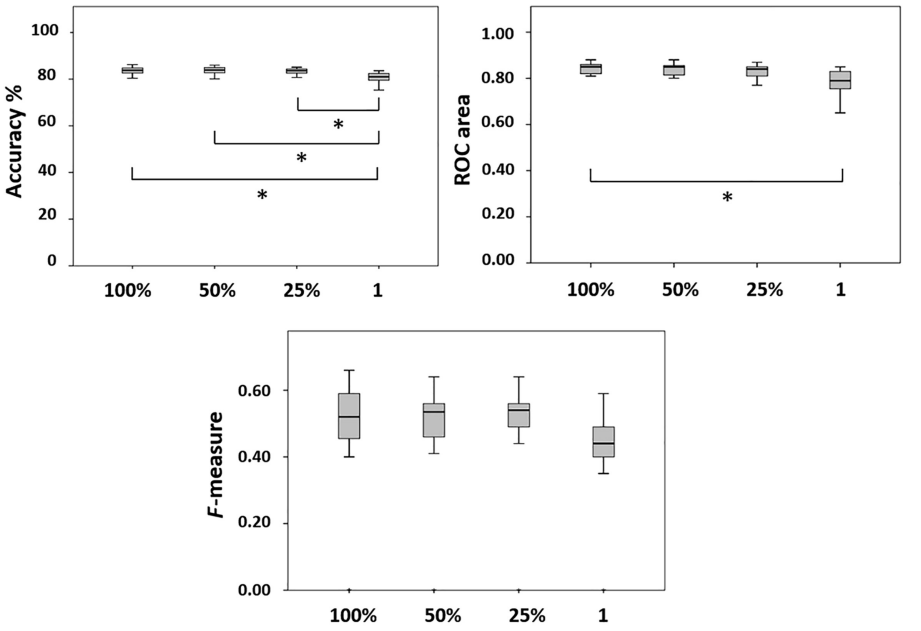

Figure 1.
Box plots, differences in accuracy (left), ROC area (in the middle) and *F*-measure (right) between each training matrix created with 100, 50 and 25% and the variable with greatest entropy

Note(s): *Statistically significant differences, $p < 0.05$

the classification model with the first variable, 25, 50 and 100% of the entropy ranking variables. For accuracy, in the *post hoc* analyses, statistically significant differences were found in each comparison with the first variable in the entropy ranking ($p < 0.001$). For the ROC value, significant differences were found between using 100% of the variable vs using the first variable of the entropy ranking ($p = 0.020$). For *F*-measure, no significant differences were found ($p = 0.050$).

The comparison in terms of accuracy, ROC value and *F*-measure, of all the classifiers constructed, trained and tested using the first variable, 25, 50 and 100% of the entropy ranking variables, are presented in Table 3. The best accuracy level, ROC value and *f*-measure were obtained with "Subspace KNN," "Medium Gaussian SVM" and "RUSboosted trees," respectively. The detail of the values obtained for the best classifiers has been highlighted in italic (Table 3).

Table 4 presents the results of the performance of the classifiers using 100% of the variables studied in their construction. These results are presented based on the predictive measures used in this study. The classifier "RUSboosted tress" yielded the lowest number of FN and the greatest sensitivity of the algorithms used. "Fine Gaussian SVM" and "Coarse KNN" achieved maximum specificity. The values obtained have been highlighted (Table 4).

Discussion

We implemented a method to predict SA in the upper years of a physiotherapy program using machine learning with the university's data. The main result obtained in this study indicates that the most sensitive model to predict attrition is "RUSboostedtrees," using 100% of the analyzed variables. This classifier's success could be explained by the asymmetry between the attrition and non-attrition groups. In this sense, previous studies have shown good accuracy in predicting attrition with decision trees (Cuji *et al.*, 2017; Kuna *et al.*, 2010; Méndez,

Classifier		Using 100% of the variables	Using 50% of the variables with greater entropy	Using 25% of the variables with greater entropy	Using the first variable in the entropy ranking
Complex tree	Accuracy	84.8%	83.0%	84.2%	81.0%
	ROC area	0.75	0.81	0.81	0.74
	F-measure	0.52	0.53	0.55	0.50
Medium tree	Accuracy	84.8%	82.7%	84.8%	81.1%
	ROC area	0.74	0.81	0.81	0.74
	F-measure	0.50	0.52	0.58	0.5
Simple tree	Accuracy	83.8%	85.1%	85.1%	83.6%
	ROC area	0.81	0.82	0.82	0.82
	F-measure	0.55	0.62	0.62	0.63
Linear discriminant	Accuracy	85.3%	86.0%	84.5%	83.0%
	ROC area	0.85	0.85	0.85	0.85
	F-measure	0.64	0.64	0.61	0.44
Quadratic discriminant	Accuracy	84.3%	85.4%	84.8%	83.0%
	ROC area	0.82	0.83	0.85	0.84
	F-measure	0.61	0.64	0.62	0.44
Logistic regression	Accuracy	82.4%	84.5%	85.1%	82.7%
	ROC area	0.85	0.87	0.86	0.85
	F-measure	0.59	0.55	0.54	0.38
Linear SVM	Accuracy	84.3%	85.4%	83.9%	82.1%
	ROC area	0.84	0.86	0.86	0.85
	F-measure	0.52	0.55	0.50	0.35
Quadratic SVM	Accuracy	80.4%	84.8%	83.0%	83.0%
	ROC area	0.82	0.81	0.82	0.81
	F-measure	0.59	0.56	0.51	0.47
Cubic SVM	Accuracy	84.3%	85.7%	82.1%	30.7%
	ROC area	0.85	0.8	0.72	0.33
	F-measure	0.63	0.61	0.49	0.19
Fine Gaussian SVM	Accuracy	82.4%	80.1%	81.3%	82.1%
	ROC area	0.74	0.81	0.79	0.78
	F-measure	0.00	0.08	0.29	0.54
Medium Gaussian SVM	Accuracy	*82.4%*	85.7%	84.8%	81.0%
	ROC area	*0.88*	0.88	0.85	0.79
	F-measure	*0.48*	0.54	0.56	0.44
Coarse Gaussian SVM	Accuracy	82.4%	82.1%	82.1%	82.1%
	ROC area	0.85	0.86	0.86	0.84
	F-measure	0.11	0.29	0.35	0.35
Fine KNN	Accuracy	83.8%	81.0%	80.7%	75.9%
	ROC area	0.68	0.69	0.68	0.67
	F-measure	0.50	0.50	0.48	0.46
Medium KNN	Accuracy	82.8%	83.6%	84.2%	80.7%
	ROC area	0.87	0.85	0.87	0.81
	F-measure	0.41	0.41	0.52	0.44
Coarse KNN	Accuracy	82.4%	80.7%	80.7%	80.7%
	ROC area	0.83	0.84	0.85	0.82
	F-measure	0.00	0.00	0.00	0.00
Cosine KNN	Accuracy	85.3%	83.9%	83.6%	77.7%
	ROC area	0.86	0.85	0.85	0.78
	F-measure	0.53	0.48	0.54	0.59
Cubic KNN	Accuracy	82.8%	83.9%	83.6%	80.7%
	ROC area	0.86	0.85	0.86	0.81
	F-measure	0.40	0.43	0.52	0.44

(*continued*)

Table 3.
Accuracy of the classifiers according to variables used

Classifier		Using 100% of the variables	Using 50% of the variables with greater entropy	Using 25% of the variables with greater entropy	Using the first variable in the entropy ranking
Weighted KNN	Accuracy	84.8%	82.7%	83.3%	78.9%
	ROC area	0.87	0.86	0.84	0.72
	F-measure	0.51	0.47	0.54	0.39
Boosted trees	Accuracy	83.3%	84.2%	83.0 %	80.1%
	ROC area	0.85	0.83	0.81	0.78
	F-measure	0.58	0.56	0.54	0.48
Bagged trees	Accuracy	85.8%	84.8%	83.3%	78.3%
	ROC area	0.87	0.85	0.81	0.77
	F-measure	0.59	0.56	0.54	0.41
Subspace discriminant	Accuracy	85.8%	83.9%	83.0%	83.0%
	ROC area	0.85	0.86	0.85	0.85
	F-measure	0.51	0.46	0.44	0.44
Subspace KNN	Accuracy	*86.3%*	83.0%	80.7%	75.3%
	ROC area	*0.85*	0.82	0.77	0.65
	F-measure	*0.43*	0.44	0.49	0.43
RUSboosted trees	Accuracy	*83.3%*	82.4%	83.9%	81.5%
	ROC area	*0.87*	0.85	0.81	0.79
	F-measure	*0.66*	0.64	0.64	0.59

Note(s): Percentage accuracy values, ROC area and *F*-measure of the classifier used; SVM = support vector machine; KNN = *k*-nearest neighbors; The best classifiers has been highlighted in italic

Table 3.

Classifier	TN	FP	FN	TP	Sen	Spec	PPV	NPV
Complex tree	239	32	31	34	52%	88%	52%	89%
Medium tree	238	33	32	33	51%	88%	50%	88%
Simple tree	248	23	32	33	51%	92%	59%	89%
Linear discriminant	247	24	23	42	65%	91%	64%	91%
Quadratic discriminant	240	31	23	42	65%	89%	58%	91%
Logistic regression	258	13	32	33	51%	95%	72%	89%
Linear SVM	256	15	37	28	43%	94%	65%	87%
Quadratic SVM	253	18	30	35	54%	93%	66%	89%
Cubic SVM	252	19	26	39	60%	93%	67%	91%
Fine Gaussian SVM	271	0	65	0	0%	*100%*	Indefinite	81%
Medium Gaussian SVM	263	8	42	23	35%	97%	74%	86%
Coarse Gaussian SVM	268	3	61	4	6%	99%	57%	81%
Fine KNN	252	19	37	28	43%	93%	60%	87%
Medium KNN	267	4	47	18	28%	99%	82%	85%
Coarse KNN	271	0	65	0	0%	*100%*	Indefinite	81%
Cosine KNN	263	8	39	26	40%	97%	76%	87%
Cubic KNN	267	4	48	17	26%	99%	81%	85%
Weighted KNN	265	6	41	24	37%	98%	80%	87%
Boosted trees	245	26	28	37	57%	90%	59%	90%
Bagged trees	258	13	32	33	51%	95%	72%	89%
Subspace discriminant	262	9	40	25	38%	97%	74%	87%
Subspace KNN	262	9	45	20	31%	97%	69%	85%
RUSboosted trees	232	39	*14*	51	*78%*	86%	57%	94%

Note(s): Sen = sensitivity; Spec = specificity; TN = true negative; FP = false positive; FN = false negative; TP = true positive; PPV = positive predictive value; NPV = negative predictive value; The best classifiers has been highlighted in italic

Table 4.
Confusion matrix and predictive value by classifier

2016; Miranda and Guzmán, 2017; Moseley and Mead, 2008), logistic regression (Aulck *et al.*, 2016; Hernández *et al.*, 2016) or neural networks (Hernández *et al.*, 2016; Miranda and Guzmán, 2017). However, these classifiers first need to use a class balance algorithm (Méndez, 2016). Because of the nature of decision trees, these models are useful in classification problems (Miranda and Guzmán, 2017); adding the "RUSboost" algorithm alleviates the issue of class imbalance (Kesikoglu *et al.*, 2016; Seiffert *et al.*, 2010). Considering that the proportion of attrition student is considerably smaller, the use of "RUSboosted trees" seems interesting.

We found statistically significant differences according to the number of variables included in the training of the classifiers ($p < 0.001$). The accuracy of the classifiers increased according to the number of characteristics or variables studied. This supports the idea that the attrition issue is multivariate (Cuji *et al.*, 2017; Patiño and Cardona, 2012) and could not be explained by studying its components separately. The average accumulated grades is one factor that best explains attrition (Cuji *et al.*, 2017; Méndez, 2016). Our results indicate that this variable by itself would generate a classifier with 83.6% accuracy using "Simple tree". In addition, two of the 23 classifiers studied demonstrated greater accuracy when only this variable was used. Despite this, our results indicate a significantly lower accuracy ($p < 0.001$) and lower ROC area ($p = 0.020$) of the classifiers using only this variable.

Grade-point average has been related to academic success (Bowers, 2010; Young *et al.*, 2011). Furthermore, there are significant differences in this variable between the attrition and non-attrition group. However, we found discrepancies in the literature regarding whether this variable alone might not be enough to differentiate attrition from non-attrition (Vásquez and Miranda, 2019). This could be explained because attrition is a complex phenomenon, involving family aspects, vocational problems, financial and academic issues, as well as other factors such as the policies and practices of the higher education institutions themselves (Reay, 2004; Tarabini and Ingram, 2015; Thomas, 2002). Our results would reinforce the idea that, although grade-point average is useful and contributes to the model, the prediction of attrition is even better when combined with other sociodemographic or academic variables.

The best accuracy level was obtained by the classifier "Subspace KNN"; however, it presented lower sensitivity. "Fine Gaussian SVM" and "Coarse KNN" achieved maximum specificity. Without limiting the foregoing, the interest of the classification problem presented should be, based on the data obtained, to focus on the detection of the attrition and therefore on the sensitivity (Bravo and Cruz, 2015). Our results indicate that "RUSboosted trees" obtained the greatest sensitivity (80.56%) of the classifiers studied; therefore, it is the best suited to detecting SA.

The strengths of this study include the analysis being applied in a context of upper-year students. Most research into university attrition concentrates on the first year of entry (Canales and De los Ríos, 2007; Himmel, 2002), this being a key level for attrition (Silva, 2011). However, there is an even larger proportion of attrition students in upper-year courses, and that phenomenon has scarcely been studied (González and Uribe, 2018). With respect to the limitations of the study, an imbalanced database was used. This imbalance is explained because the proportion of attrition students is much smaller. Nevertheless, a classifier highly sensitive to attrition was identified. The selected characteristics or variables are not sufficient to explain the problem in its entirety. Furthermore, and due to the sample size, our results are difficult to generalize.

With respect to the practical connotations of this study, first, the predictive model identified could be used in databases of other universities, if the respective local reality and the registration of similar variables are previously considered. In this sense, we emphasize that a method to identify SA risk was established using databases with a significant class imbalance, in this case, with a considerably smaller proportion of attrition students. Secondly, at the local level, the implementation of this proposal will make available a listing that will contain the dichotomous classification of the attrition risk for each student, which would

facilitate early detection and intervention using the university's devices. This would contribute to make methodological adjustments on the part of the academics. As future projections, on the one hand, it is expected that the training of the classifier can be replicated with new cohorts of students, using databases from other programs, international free-access databases and databases from partner universities, thus standardizing the use of this tool and seeing whether the accuracy of the classifiers studied can be improved. On the other hand, as we have pointed out previously, this system was implemented in a university's physiotherapy program to study if it is a useful measurement in decreasing attrition levels locally.

In conclusion, the use of a classifier based on supervised machine learning "RUSboosted trees" made it possible to identify upper-year attrition students from a physiotherapy program, using student characterization records and the university's own databases. In addition, it proved useful when the proportion of attrition students was considerably smaller, and its implementation in similar contexts could be considered.

References

Adejo, W. and Connolly, T. (2018), "Predicting student academic performance using multi-model heterogeneous ensemble approach", *Journal of Applied Research in Higher Education*, Vol. 10 No. 1, pp. 61-75, doi: 10.1108/JARHE-09-2017-0113.

Agarwal, S., Pandey, G.N. and Tiwari, M.D. (2012), "Data mining in education: data classification and decision tree approach", *International Journal of E-Education, e-Business, e-Management and e-Learning*, Vol. 2 No. 2, pp. 140-145.

Amaya, K., Barrientos, E. and Heredia, J. (2014), "Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos", *Mining Techniques*, Vol. 13 No. 9, pp. 3127-3134, available at: http://repositorio.redclara.net/handle/10786/759.

Asociación Médica Mundial (2000), "Declaración de Helsinki de la AMM – Principios éticos para las investigaciones médicas en seres humanos – WMA – The World Medical Association", available at: https://www.wma.net/es/policies-post/declaracion-de-helsinki-de-la-amm-principios-eticos-para-las-investigaciones-medicas-en-seres-humanos/ (accessed 1 October 2018).

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. (2016), "Predicting student dropout in higher education", in *ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, New York, doi: 10.1002/prot.24187.

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zárate-Valderrama, J. and Yanque-Churo, P. (2020), "Classification models for determining types of academic risk and predicting dropout in university students", *International Journal of Advanced Computer Science and Applications*, Vol. 11 No. 1, doi: 10.14569/IJACSA.2020.0110133.

Bowers, A.J. (2010), "Grades and graduation: a longitudinal risk perspective to identify student dropouts", *The Journal of Educational Research*, Vol. 103 No. 3, pp. 191-207, doi: 10.1080/00220670903382970.

Bravo, S. and Cruz, J. (2015), "Estudios de exactitud diagnóstica: Herramientas para su Interpretación Diagnostic accuracy studies: tools for its interpretation", *Revista Chilena de Radiología. Año*, Vol. 21 No. 4, pp. 158-164, doi: 10.4067/S0717-93082015000400007.

Braxton, J.M. (2019), "Leaving college: rethinking the causes and cures of student attrition by Vincent Tinto", *Journal of College Student Development*, Vol. 60 No. 1, pp. 129-134, doi: 10.1353/csd.2019.0012.

Bruneforth, M., Motivans, A. and Zhang, Y. (2004), *Investing in the Future: Financing the Expansion of Educational Opportunity in Latin America and the Caribbean*, I. de estadística de la UNESCO (Ed.), Unesco, Montreal, available at: http://uis.unesco.org/sites/default/files/documents/investing-in-the-future-financing-the-expansion-of-educational-opportunity-in-latin-america-and-the-caribbean-04-en_0.pdf.

Canales, A. and De los Ríos, D. (2007), "Factores explicativos de la deserción universitaria", *Calidad En La Educación*, Vol. 0 No. 26, pp. 173-201, doi: 10.31619/caledu.n26.239.

Chen, Y., Johri, A. and Rangwala, H. (2018), "Running out of STEM", *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, New York, NY, ACM, pp. 270-279. doi: 10.1145/3170358.3170410.

CNED (2020), "Consejo Nacional de Educación – Chile. Indicadores de duración y retención en Educación Superior, años 2014–2019", available at: https://www.cned.cl/indices/duracion-y-retencion-anos-2014-2019 (accessed 3 April 2021).

Costa, E.B., Fonseca, B., Santana, M.A., de Araújo, F.F. and Rego, J. (2017), "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses", *Computers in Human Behavior*, Vol. 73, pp. 247-256, doi: 10.1016/j.chb.2017.01.047.

Cuji, B., Gavilanes, W. and Sanchez, R. (2017), "Modelo predictivo de deserción estudiantil basado en arboles de decisión", *Espacios*, Vol. 38 No. 55, p. 17.

Dutt, A., Aghabozrgi, S., Akmal, M., Ismail, B. and Mahroeian, H. (2015), "Clustering algorithms applied in educational data mining", *International Journal of Information and Electronics Engineering*, Vol. 5 No. 2, pp. 112-116, doi: 10.7763/IJIEE.2015.V5.513.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "From data mining to knowledge discovery in databases", *AI Magazine*, Vol. 17 No. 3, p. 37, doi: 10.1609/AIMAG.V17I3.1230.

Frank, E., Hall, M.A. and Witten, I.H. (2016), *The WEKA Workbench*, 4th ed., Morgan Kaufmann, Burlington, MA, pp. 553-571. doi: 10.1016/B978-0-12-804291-5.00024-6.

Gallegos, J.A., Campos, N.A., Canales, K.A. and González, E.N. (2018), "Factores Determinantes en la Deserción Universitaria. Caso Facultad de Ciencias Económicas y Administrativas de la Universidad Católica de la Santísima Concepción (Chile)", *Formación Universitaria*, Vol. 11 No. 3, pp. 11-18, doi: 10.4067/s0718-50062018000300011.

Gil, J.S., Delima, A.J.P. and Vilchez, R.N. (2020), "Predicting students' dropout indicators in public school using data mining approaches", *International Journal of Advanced Trends in Computer Science and Engineering*, Vol. 9 No. 1, pp. 774-778, doi: 10.30534/ijatcse/2020/110912020.

González, L.E. and Uribe, D. (2018), "Estimaciones sobre la 'repitencia' y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones", *Calidad En La Educación*, Vol. 0 No. 17, p. 75, doi: 10.31619/caledu.n17.408.

Harrington, P. (2012), *Machine Learning in Action: Examples. Machine Learning in Action: Examples*, Manning Publications, New York, doi: 10.1007/978-1-4302-5990-9_11.

Hernández, A.G., Meléndez, R.A., Morales, L.A., Garcia, A., Tecpanecatl, J.L. and Algredo, I. (2016), "Comparative study of algorithms to predict the desertion in the students at the ITSM-Mexico", *IEEE Latin America Transactions*, Vol. 14 No. 11, pp. 4573-4578, doi: 10.1109/TLA.2016.7795831.

Himmel, E. (2002), "Modelo de análisis de la deserción estudiantil en la educación superior", *Calidad En La Educación*, Vol. 0 No. 17, p. 91, doi: 10.31619/caledu.n17.409.

Kesikoglu, M.H., Atasever, U.H., Ozkan, C. and Faculty, E. (2016), "The usage of RUSBoost boosting method for classification of impervious surfaces", *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLI, pp. 981-985, doi: 10.5194/isprs-archives-XLI-B7-981-2016.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers, Montreal, pp. 1137-1143.

Kuna, H., García, R. and Villatoro, F. (2010), "Pattern discovery in university students desertion based on data mining", *Proceedings of The IV Meeting on Dynamics of Social and Economic Systems*, Vol. 2 No. 2, p. 11, available at: http://iidia.com.ar/rgm/articulos/AASSJ-4091023-SD48-Kuna.pdf.

Mathworks (2021), "Ensemble algorithms", available at: https://la.mathworks.com/help/stats/ensemble-algorithms.html#mw_0d10d11d-5e75-477d-8957-18933764faf3.

Mduma, N., Kalegele, K. and Machuve, D. (2019), "A survey of machine learning approaches and techniques for student dropout prediction", *Data Science Journal*, Vol. 18, pp. 1-10, doi: 10.5334/dsj-2019-014.

Méndez, J.J. (2016), "Proyección de Estudiantes en Riesgo de Desertar Mediante Técnicas de Minería de Datos", *Ingeniería, Innovación y Desarrollo Sostenible*, Vol. 1 No. 1, pp. 23-35, doi: 10.21892/25008803.179.

Miranda, M.A. and Guzmán, J. (2017), "Análisis de la deserción de estudiantes universitarios usando técnicas de minería de datos", *Formacion Universitaria*, Vol. 10 No. 3, pp. 61-68, doi: 10.4067/S0718-50062017000300007.

Moseley, L.G. and Mead, D.M. (2008), "Predicting who will drop out of nursing courses: a machine learning exercise", *Nurse Education Today*, Vol. 28 No. 4, pp. 469-475, doi: 10.1016/j.nedt.2007.07.012.

Murakami, K., Takamatsu, K., Kozaki, Y., Kishida, K., Bannaka, K., Noda, I., Asashi, J., Takao, K., Mitsunari, K., Nakamura, T. and Nakata, Y. (2019), "Predicting the probability of student dropout through EMIR using data from current and graduate students", *Proceedings – 2018 7th International Congress on Advanced Applied Informatics, IIAI-AAI*, pp. 478-481, doi: 10.1109/IIAI-AAI.2018.00103.

Ortiz-Lozano, J.M., Rua-Vieites, A., Bilbao-Calabuig, P. and Casadesús-Fa, M. (2018), "University student retention: best time and data to identify undergraduate students at risk of dropout", *Innovations in Education and Teaching International*, Vol. 57 No. 1, pp. 74-85, doi: 10.1080/14703297.2018.1502090.

Paluszek, M. and Thomas, S. (2019), *MATLAB Machine Learning Recipes: A Problem-Solution Approach*, 2nd ed., Apress, Berkeley, CA, New York City, doi: 10.1007/978-1-4842-3916-2.

Patiño, L. and Cardona, A.M. (2012), "Revisión De Algunos Estudios Sobre La Deserción Estudiantil Universitaria En Colombia Y Latinoamérica", *Theoria*, Vol. 21 No. 1, pp. 9-20, available at: http://www.redalyc.org/articulo.oa?id=29931769002.

Reay, D. (2004), "It's all becoming a habitus': beyond the habitual use of habitus in educational research", *British Journal of Sociology of Education*, Vol. 25 No. 4, pp. 431-444, doi: 10.1080/0142569042000236934.

Sandoval, A., Gonzalez, C., Alarcon, R., Pichara, K. and Montenegro, M. (2018), "Centralized student performance prediction in large courses based on low-cost variables in an institutional context", *Internet and Higher Education*, Vol. 37, pp. 76-89, doi: 10.1016/j.iheduc.2018.02.002.

Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J. and Napolitano, A. (2010), "RUSBoost: a hybrid approach to alleviating class imbalance", *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, Vol. 40 No. 1, pp. 185-197.

SIES (2020), "Sistema de Información de Educación superior – MINEDUC Chile", *Informe de Retención de primer año de Pregrado*, available at: https://www.mifuturo.cl/wp-content/uploads/2020/12/Informe_retencion_pregrado_SIES_2020.pdf (accessed 3 April 2021).

Silva, M. (2011), "Primer año universitario como un tramo crítico para el éxito académico", *Revista Perfiles Educativos*, Vol. XXXIII, pp. 57-66.

Siri, A. (2015), "Predicting students' dropout at university using artificial neural networks", *Italian Journal of Sociology of Education*, Vol. 7 No. 2, pp. 225-247.

Stinebrickner, T. and Stinebrickner, R. (2007), *The Effect of Credit Constraints on the College Drop-Out Decision A Direct Approach Using a New Panel Study*, National Bureau of Economic Research, Cambridge, Massachusetts, doi: 10.3386/w13340.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. and Abreu, R. (2015), "A comparative study of classification and regression algorithms for modelling students' academic performance",

Proceedings of the 8th International Conference on Educational Data Mining, pp. 392-395, available at: http://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf.

Tarabini, A. and Ingram, N. (2015), *Educational Choices, Transitions and Aspirations in Europe*, Routledge, London.

Thomas, L. (2002), "Student retention in higher education: the role of institutional habitus", *Journal of Education Policy*, Vol. 17 No. 4, pp. 423-442, doi: 10.1080/02680930210140257.

Tinto, V. (1993), "Reflexiones sobre el abandono de los estudios superiores", *Perfiles Educativos*, Vol. 62 No. 62, pp. 56-63, available at: http://132.248.192.201/seccion/perfiles/1993/n62a1993/mx.peredu.1993.n62.p56-63.pdf.

Vásquez, J. and Miranda, J. (2019), "Student desertion: what is and how can it be detected on time?", in *Data Science and Digital Business*, Springer International Publishing, Cham, pp. 263-283, doi: 10.1007/978-3-319-95651-0_13.

Villamizar, G.A. and Romero, M.L. (2011), "Relación entre variables psicosociales y rendimiento académico en estudiantes de primer semestre de psicología", *Educación y Desarrollo Social*, Vol. 5 No. 1, pp. 41-54, doi: 10.18359/reds.891.

Viloria, A., Padilla, J.G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N.O. and David, M.A. (2019), "Integration of data technology for analyzing university dropout", *Procedia Computer Science*, Vol. 155, pp. 569-574, doi: 10.1016/j.procs.2019.08.079.

Wan Yaacob, W.F., Mohd Sobri, N., Nasir, S.A.M., Wan Yaacob, W.F., Norshahidi, N.D. and Wan Husin, W.Z. (2020), "Predicting student drop-out in higher institution using data mining techniques", *Journal of Physics: Conference Series*, Vol. 1496, 012005, doi: 10.1088/1742-6596/1496/1/012005.

Young, A.E., Worrell, F.C. and Gabelko, N.H. (2011), "Predictors of success in accelerated and enrichment summer mathematics courses for academically talented adolescents", *Journal of Advanced Academics*, Vol. 22 No. 4, pp. 558-577, doi: 10.1177/1932202X11413886.

About the authors

Mauricio Barramuño is a Professor at the Faculty of Health Sciences of the Autonomous University of Chile. He works in the motion analysis laboratory associated with the university's kinesiology program. The interest and experience in research lie in the study of the human movement and its teaching through the acquisition and processing of biological data. Mauricio Barramuño is the corresponding author and can be contacted at: mauricio.barramuno@uautonoma.cl

Claudia Meza-Narváez currently serves as the Secretary of Studies and Teacher in the Kinesiology Degree at the Autonomous University of Chile. She is Master in Teaching and Pedagogical Innovation in Higher Education. In research, she is interested in the diagnostic study of the educational environment and retention.

Dr. Germán Gálvez-García is an Associate Professor of Psychology at University of La Frontera of Chile and associate member of Université Lumiere Lyon 2. He teaches undergraduate, magister and doctorate in subjects of experimental psychology and neuroscience. He is the Director of the Laboratory of Cognitive Neuroscience and Action, having published mostly within this subject.