Perspectives to Predict Dropout in University Students with Machine Learning

Martín Solís
Tecnológico de Costa Rica
marsolis@itcr.ac.cr

Tania Moreira
Tecnológico de Costa Rica
tmoreira@itcr.ac.cr

Roberto González
Tecnológico de Costa Rica
gonzalezloaizar@gmail.com

Tatiana Fernández
Tecnológico de Costa Rica
tafema@itcr.ac.

María Hernández
Tecnológico de Costa Rica
thernandez@itcr.ac.cr

Abstract— **This study analyzes the performance of four machine learning algorithms with different perspectives for defining data files, in the prediction of university student desertion. The algorithms used were: Random Forest, Neural Networks, Support Vector Machines and Logistic Regression. It was found that the Random Forest algorithm with 10 variables randomly sampled as candidates in each division, was the best for predicting dropouts and that the ideal perspective for training the algorithm is to use information on all semesters that students take within a given period of time, using a classification variable that defines the non-dropout as the graduated student. In a first validation sample, this approach correctly predicted 91% of dropouts, with a sensitivity of 87%**.

Keywords—Dropout, university students, machine learning.

I. INTRODUCTION

It is imperative to analyze the phenomenon of desertion in public higher education. According to [1], dropout rates and low graduation rates have become a matter of increasing interest for higher education institutions and education authorities over the last fifty years, given that low graduation rates increase social and economic gaps between social groups, and hinder the development of countries. This has led to the development of research that is looking for algorithms for predicting university dropout rates, whose results can guide the design of support programs that will increase student retention.

The most recent research in this area have focused on accurately predicting the risk of students dropping out during their first semester of university studies. These research have analyzed the phenomenon over different periods, with some focusing on only students first year of studies, while others cover up to 5 or more years of university studies [3, 4]. Likewise, the criteria used for defining dropout have varied, from not enrolling in the first or second semester of the first year of studies [5, 6, 7] to longer periods, equivalent to 3 years during which they do not enroll [4].

Several machine learning algorithms have been used in these studies, including decision trees, Naïve Bayes [7]; neural networks [6]; logistic regression [8], and Bayesian networks [4]. In general, these techniques allow the identification of patterns and associations between the variables included in the algorithms, and dropping out [9]. As emphasized in [8], a static approach to identifying students at risk is not advisable; it is therefore necessary to use dynamic models and test these using data from several semesters.

Several types of variables have been used in this type of research as predictors of dropout, such as those related to academic, financial and sociodemographic characteristics [2], scores obtained in university entrance test [7], place of student residence [10, 11], secondary school of student origin [12], and academic programs of study [11]. The combination of these variables has made it possible to improve the sensitivity analysis of predictive models, and has provided important information about individual and institutional factors that increase the probability of dropping out [2]. Therefore, in this research it is included several types of variables, as such sociodemographic, program of study, academic history and semesters recorded. This information is organized in the Institutional Management Indicator System (SIGI) of Institutional Planning Office.

The present investigation, just as those mentioned previously, uses several machine learning algorithms to train an algorithm to predict university student dropouts at the Instituto Tecnológico de Costa Rica. The principal novel contribution of this investigation is that four perspectives are analyzed on how the data file should be defined for training and validation of the algorithms. The perspectives arise from combining two elements:

1. The way in which the non-dropout is defined. One issue when trying to predict future dropouts is whether active students should be included in the group of non-dropouts. Including them can add noise

to the training and prediction of the algorithm, since it is not known if they will stop enrolling or they will graduate in the future. An alternative is to use only those who have graduated as non-dropouts, excluding active students. Both of these approaches are analyzed in this study.

2. The choice of the semester. This refers to whether data from all semesters in which dropouts and non-dropouts are enrolled in a given period are included in the analysis, or only last enrollment before dropping out, which provides the most updated information for the prediction. Another contribution of the present investigation is that most studies have focused on predicting those who will drop out after the first semester or first academic year [2,3,5,6, 8,12,13], while the present research attempts to predict who will stop enrolling in the future. The objective is to train an algorithm that can be used at the end of each semester (not only for the first ones), to determine which students will drop out (stop enrolling for at least two consecutive semesters in the future), in order to take steps to address these dropouts before they occur.

II. Method

A. Data

The sample is composed of all those students who enrolled in a degree program at the Instituto Tecnológico de Costa Rica (ITCR) between the years 2011 and 2016. There were 90,067 records, corresponding to the enrollments of 16,807 students, which initially met this criterion. Records that were incomplete or contained incorrect information were deleted, resulting in a final sample of 80,527 records of 15,720 students. Four perspectives were used to predict dropouts, each of which involved the analysis of different numbers of records:

1. The first perspective uses data from all records for students who enrolled in the semesters between the years 2011 to 2016. A dropout is considered to be a student who has at least two academic years without enrolling and who has not yet graduated, while a non-dropout is an active student or one who finished his or her studies between 2011 and 2016. A total of 80,527 records satisfied these criteria, of which 19.1% belonged to students who were classified as dropouts.
2. In the second perspective a dropout is defined in the same way as in the previous perspective; however, a non-dropout is defined as a student who finished his or her studies. The objective here is to eliminate noise of active students when training the algorithm, since it is not known beforehand if they will graduate or abandon. A total of 35,132 records satisfied these criteria, of which 43.7% belonged to students who were defined as dropouts.
3. In the third perspective, dropouts and non-dropouts are defined as in perspective 1. The difference is that only one semester (one period) is used for each student who enrolled between 2011 and 2016. In the case of dropouts, only information of the last semester before dropping out is used, and in the case of non-dropouts, a semester is chosen at random. The purpose of this perspective is to eliminate noise from previous semesters of the dropout, on the assumption that the most recent semester provides the most up-to-date information to predict if he or she is going to drop out. A total of 15,720 records satisfied these criteria, of which 28% belong to students who were defined as dropouts.
4. In this perspective, the definition of dropouts and non-dropouts is the same as that used in perspective 2, but only one semester per student is used, as was done in third perspective. The objective of this perspective is to eliminate noise from active students and from previous semesters of students who drop out. A total of 7,936 records satisfied these criteria, of which 55.7% belong to students who were defined as dropouts.

B. Variables

There are two variables that are used to define a "dropout". In the first variable, a student who has passed at least two academic years without enrolling and who has not graduated is considered a dropout, while the non-dropout is a student who has graduated or is still active. In the second variable, the definition of a dropout is the same used in the first variable, while a non-dropout is defined as a student who graduated. Table 1 shows these dropout variables and other variables that were used as predictors. There are a total of 19 variables, which can be classified into three groups.

Table 1. Variable definitions and types of measurement

Variable	Measurement
Dropout 1. Student who has at least two semesters without enrolling and who has not yet graduated	0 = Non-dropout (active or graduate) 1 = Dropout
Dropout 2	0 = Graduated 1 = Dropout
Year of enrollment	Date
Sociodemographic	
Gender	0 = Male, 1 = Female
Residence while attending classes	0 = Same during classes and vacation 1 = Different when not attending classes
Received the highest	0 = No, 1 = Yes

scholarship of the institution during the semester	
Received a financial grant (loan) during the semester	0 = No, 1 = Yes
Program	
Type of program in which the student is enrolled	0 = Bachelor 4 years 1 = Bachelor 5 years
Location where the student is enrolled	1 = Alajuela 2 = Cartago … 5 = San José
Study program in which the student is enrolled	1 = Agro Business Engineering 2 = Business Administration 3 = Agronomic Engineering … 23 = Electrical Engineering
School shift	0 = Nighttime, 1 = Daytime
Student admitted to first choice of career	0 = No, 1 = Yes
Student is participating in first choice if career.	0 = No, 1 = Yes
Student requested change of career.	0 = Not requested 1 = Requested and accepted 2 = Requested and rejected
Academic history	
Grade average for the semester	Grade from 0 to 100
Annual academic cycle (semester)	0 = First, 1 = Second
Courses taken in the semester	Quantity
Courses approved in the semester	Quantity
Semesters not enrolled in the past	Quantity
Year of admission to the university	Academic year
Courses needed to graduate	Quantity

C. Procedure

The ability of four types of algorithms to predict dropping out was evaluated: Random Forest, Neural Networks Multilayer Perceptron, Support Vector Machines with a Radial Basis Function (RBF) and Logistic Regression. To determine the predictive capacity of the variables and choose the best prediction method, the following steps were carried out:

1. Dividing the data: For predictive purposes the data must be divided into at least two parts [14], one to train the model and the other to evaluate its predictive capacity. The first part was made up of data for 75% of the subjects, chosen at random, and the second part by the remaining 25%.

2. Training: The parameters of the four methods were estimated in this phase. For this process the Caret library (Classification and Regression Training) of the R software was used. Using this library, the models are calibrated by evaluating different parameters in each method by default [15]. To determine which parameters contribute to better prediction, a 5-fold cross-validation approach was used, since it minimizes the variance associated with the validation process [14]. This procedure consists of taking 80% of the sample to estimate the algorithm with specific parameters and 20% to classify the observations (5 non-overlapping training and testing sets). This is replicated five times. At the end of the process, the classification results of these five replications are averaged. The predictive capacity of the algorithm is evaluated with the Kappa coefficient shown below.

$$k = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (1)$$

Where Pr refers to the proportion of real (a) and expected (e) agreement between the classifier and the true values. A Kappa value greater than 0.60 is considered a good fit [14].

For each algorithm, 3 different values of each of the parameters were tested. At the end those parameters that generated the highest Kappa value were selected.

In Random Forest the parameter tested was mtry. This is the number of variables randomly sampled as candidates in each division. In Support Vector Machines the parameters tested was C and Sigma. C is used for the soft margin cost function (involves trading error penalty for stability) while Sigma is the standard deviation. In Neuronal Network the parameters tested were size and decay. The first one represents de number of hidden layers and the second is the value of the weight update rule that causes the weights to exponentially decay to zero. By last, the parameter tested in logistic regression were decay, which specify the norm used in the penalization.

3. Validation: In this step, the predictive capacity of the six estimated models was evaluated. To do so, the predictions of the six models were initially obtained using the 25% of the sample that had previously been set aside. Subsequently, adjustment indicators were estimated: the probability of correctly detecting dropouts, the probability of correctly detecting non-dropouts, sensitivity, specificity, and the Kappa coefficient. The probability of correctly detecting non-dropouts and the specificity are included simply for informative purposes, since the best model is the one that simultaneously maximizes the probability of correctly detecting dropouts and sensitivity.

III. RESULTS

Tables 2, 3, 4 and 5 present the performance metrics for the cross-validation of the four algorithms with the parameters that generated the highest kappa value in each of the four perspectives. In Table 2, it can be seen that under perspective 1 the sensitivity and kappa values in the four algorithms are low. The values in Table 3 show that the kappa, sensitivity, and true positive metrics improve considerably under perspective 2. The best algorithms are the Random Forest and Support Vector Machine. With prospects 3 and 4 the Random Forest is the best algorithm, although its performance is only slightly better than those of the others. Table 5 shows that the fourth perspective has the best performance. Using the Random Forest algorithm, a sensitivity of 93% and a percentage of true positives of 94% is obtained.

Table 2. Perspective 1. Validation performance metrics

Indicators	RF	SVM	NNET	LOGIT
Kappa	0.47	0.38	0.39	0.36
Sensitivity	0.46	0.32	0.38	0.33
Specificity	0.95	0.98	0.95	0.96
Positive	0.68	0.76	0.65	0.65
Negative	0.88	0.86	0.87	0.86

RF=Random Forest, SVM=Support Vector Machines, NNET=Neuronal Network, LOGIT=Logistic Regression
Parameters. RF: mtry = 10; SVM: sigma = 0.0118 and C = 1; NNET: size = 1 and decay = 0.1; LOGIT: decay = 0.1

Table 3. Perspective 2. Validation performance metrics

Indicators	RF	SVM	NNET	LOGIT
Kappa	0.80	0.80	0.76	0.74
Sensitivity	0.87	0.85	0.83	0.83
Specificity	0.92	0.94	0.93	0.91
Positive	0.91	0.91	0.90	0.88
Negative	0.90	0.89	0.87	0.87

RF=Random Forest, SVM=Support Vector Machines, NNET=Neuronal Network, LOGIT=Logistic Regression
RF: mtry = 10; SVM: sigma = 0.0119 and C = 1;
NNET: size = 5 and decay = 0.1; LOGIT: decay = 0.1

Table 4. Perspective 3. Validation performance metrics

Indicators	RF	SVM	NNET	LOGIT
Kappa	0.660	0.640	0.63	0.63
Sensitivity	0.720	0.670	0.69	0.68
Specificity	0.920	0.940	0.92	0.92
Positive	0.780	0.800	0.77	0.77
Negative	0.890	0.880	0.88	0.88

RF=Random Forest, SVM=Support Vector Machines, NNET=Neuronal Network, LOGIT=Logistic Regression
RF: mtry = 2; SVM: sigma = 0.0118 and C = 1;
NNET: size = 1 and decay = 0.1; LOGIT: decay = 0.1

Table 5. Perspective 4. Validation performance metrics

Indicators	RF	SVM	NNET	LOGIT
Kappa	0.85	0.840	0.84	0.84
Sensitivity	0.930	0.900	0.92	0.91
Specificity	0.930	0.940	0.92	0.93
Positive	0.940	0.950	0.94	0.95
Negative	0.900	0.890	0.9	0.89

RF=Random Forest, SVM=Support Vector Machines, NNET=Neuronal Network, LOGIT=Logistic Regression
RF: mtry = 2; SVM: sigma = 0.0118 and C = 1;
NNET: size = 1 and decay = 0.1; LOGIT: decay = 0.1

Although it can be seen in the results of perspective 4 that the best performance occurs with the "Random Forest" algorithm, it should also be noted that under each perspective the analysis units are modified in the validation sample; therefore, validation metrics are not directly comparable between perspectives. For this reason, a final test was carried out to more equitably compare the performance of this algorithm with perspectives 2, 3 and 4, which was where acceptable kappa coefficients were obtained (greater than 0.60).

For this validation test, the "Random Forest" algorithm was executed under each perspective, excluding the subjects that entered in 2010. Subsequently, for each semester since 2011, a prediction was made and the sensitivity and true positives metrics were estimated, considering only dropouts and graduates, since the active students were not yet defined. Figure 1 shows the value of these metrics by semester and perspective. When analyzing the graph, it was found that greater stability in the correct prediction of dropouts (true positives) occurs in perspective 2. The percentage of true positives starts at 74% and grows until reaching 95% in semester 12. Under the other perspectives, the probability of correct prediction begins to fall after semester 6 until reaching 52% under perspective 3 and 67% under perspective 4. The sensitivity values for perspective 2 were higher than those of the other perspectives until semester 7, after which the perspective 2 percentage continues to grow, but not as rapidly as the percentages under perspectives 3 and 4. Overall, perspective 2 also provides better performance, with 82% correct predictions of dropouts, and a sensitivity of 71%.

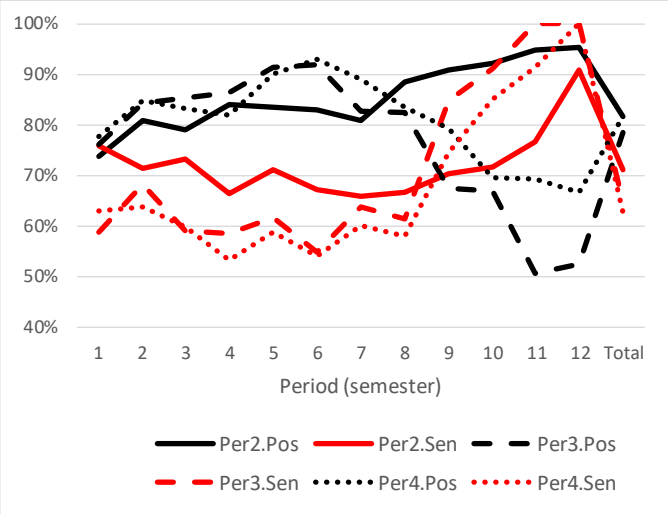

Fig 1. True positive and sensitivity by perspective and period. *Note. Per=perspective, Pos=True positive, Sen=Sensitivity*

IV. Conclusion

After analyzing the results it was determined that the best algorithm for classifying dropouts is the Random Forest with mtry = 10. The ideal perspective for building the algorithm is to use information for all of the semesters that the students enroll in, taking as a classification variable that which defines non-dropouts as those students who finished their program of study. The predictive capacity of this algorithm was the best of the alternatives evaluated. In a first validation, using 25% of the sample showed a percentage of correct detection of dropouts of 91% and a sensitivity of 87%. In a second validation, using the 2010 cohort and monitoring it over time, the percentage of correct detection of dropouts was 82% and the sensitivity was 71%. This last validation is more reliable, since a cohort is followed through time, as it would be in a real application of the algorithm.

In addition to yielding the best sensitivity metrics and true positives, the Random Forest with the perspective discussed previously shows a smaller gap between these metrics, and more appropriate behavior through time. It was found that both the ability to correctly detect dropouts and sensitivity increase over time. While using perspectives 3 and 4 it was observed that sensitivity increases, but the probability of correct detection begins to decrease from the seventh semester.

The results also suggest that, to train the dropout prediction algorithm, it is convenient to exclude active students, who may add noise because it is not known beforehand if they will dropout or graduate in the future. In essence, the problem is that they may have a dropout pattern, but they have not been classified as such.

Other predictive models of university dropout, such as those published by [2, 9] used similar methods of machine learning and achieved better sensitivity and probability of correct detection of dropouts, than that found with the second validation test; however, these research focused only on predicting those students who stopped enrolling after the first study year.

The next step to obtain an algorithm with greater precision and specificity is to include new variables in the information system, which have shown some evidence of relationship with dropping out in other studies conducted in Costa Rica [16, 17]. For example, students' interest in the programs of study to which they were admitted, the students' level of interest in studying at the university to which they were admitted, socioeconomic conditions in the students' homes, and whether or not they are working while studying. In addition, it is necessary to take variables into account that have consistently shown a strong correlation with dropping out in studies carried out in other countries, such as the educational level of the student's relatives, support from the family, and student attitudes and personalities.

References

[1] Castaño, E., Gallón, S., Gómez, K. y Vásquez, J, "Análisis de los factores asociados a la deserción y graduación estudiantil universitaria", Lecturas de Economía, no 65, pp. 9-36, 2006 .

[2] Delen, D, "A comparative analysis of machine learning techniques for student retention management", *Decision Support Systems*, vol. 49, no. 4, pp. 498-506, 2010

[3] Mesarić, J. & Šebalj, D, "Decision trees for predicting the academic success of Students", *Croatian Operational Research Review*, no.7, pp.367-388, 2016

[4] Miranda, M. A. & Guzmán, J, Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos, *Formación Universitaria*, vol. 10, no.3, pp.61-68, 2017

[5] Aguiar, E., Ambrose, G.A., Chawla, N.V., Goodrich, V & Brockman, J, "Engagement vs Performance: Using Electronic Portfolios to Predict First Semester Engineering Student Persistence", *Journal of Learning Analytics*, vol. 1, no. 3, pp. 7–33, 2014.

[6] Plagge, Mark, "Using artificial neural networks to predict first-year traditional students second year retention rates," *Proceedings of the 51st ACM Southeast Conference* ACM, 2013.

[7] Oñate, A., *Análisis de la deserción y permanencia académica en la educación superior aplicando minería de*

datos, Diss. Universidad Nacional de Colombia-Sede Bogotá, 2016

[8] Kemper, L., Gerrit, V, & Berthold, W. "Predicting Student Dropout: a Machine Learning Approach", 2017 recover from https://www.researchgate.net/profile/Lorenz_Kemper/publication/322919234_Predicting_Student_Dropout_a_Machine_Learning_Approach/links/5a7615c2a6fdccbb3c07aa70/Predicting-Student-Dropout-a-Machine-Learning-Approach.pdf

[9] Pal, S, "Mining educational data to reduce dropout rates of engineering, *International Journal of Information Engineering and Electronic Business*, vol.4, no.2,pp.1-12, 2012.

[10] Nakhkob, B. & Khademi, M, "Predicted Increase Enrollment in Higher Education Using Neural Networks and Data Mining Techniques", *Journal of Advances in Computer Research*, vol.7, no. 4, pp. 125-140, 2016.

[11] Jia, J.W, "*Machine learning algorithms and predictive models for undergraduate student retention at an HBCU"*. Diss. Bowie State University, 2013.

[12] Oztekin, A, "A hybrid data analytic approach to predict college graduation status and its determinative factors", *Industrial Management & Data Systems*, vol.116, no.8, pp.1678-1699, 2016.

[13] S. K. Yadav, B.K. Bharadwaj & S. Pal, "Data Mining Applications: A comparative study for Predicting Student's Performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19, 2011.

[14] Lantz, Brett. *Machine learning with R*. Packt Publishing Ltd, 2013.

[15] Kuhn, "M.Caret package", *Journal of statistical software*, vol. 28, no. 5, pp.1-26, 2008.

[16] Chinchilla, S, 2013, "Algunos datos sobre la deserción en el Instituto Tecnológico de Costa Rica", Informe para el Instituto Tecnológico de Costa Rica, 2013

[17] Abarca, A. & Sánchez, M. "La deserción estudiantil en la educación superior: El caso de la Universidad de Costa Rica", *Actualidades investigativas en educación*, vol.5, pp. 1-22, 2005.