Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 566–571

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop of Innovation and Technologies (IWIT 2021)

November 1-4, 2021, Leuven, Belgium

Implementation of a Predictive Information System for University Dropout Prevention

Stefania Guzmán-Castillo[a], Franziska Körner[c], Julia I. Pantoja-García[a], Lainet Nieto-Ramos[a], Yulineth Gómez-Charris[a,b], Alex Castro-Sarmiento[a], Alfonso R. Romero-Conrado[a,b]*

[a]*Universidad de la Costa, Calle 58# 55-66, Barranquilla, 08002, Colombia*
[b] *Universitat Politècnica de València, Valencia 46022, Spain*
[c]*Foris - SATD. Santiago de Chile, 8320000, Chile*

Abstract

The dropout of university students is one of the topics of a broad interest in higher education institutions and government education departments. Recent changes in education methods, socio-economic conditions, and the growing limitations of face to face interactions make it necessary to have tools that allow us to consider a broad set of factors related to the dropout phenomenon. The objective of this article is to show the implementation results of a predictive information system (IS) for the prevention of university dropout in a higher education institution. The system allows the calculation of the risk of dropout per student and uses an alert generation procedure to coordinate interventions. The platform allowed measuring the impact of the intervention strategies on the permanence of the students. Likewise, it made possible the reorganization of the intervention process towards the students, prioritizing according to the risk level.

Keywords: Dropout, prevention, education, university.

* Corresponding author. Tel.: +57-315-353-49-41.
E-mail address: aromero17@cuc.edu.co

10.1016/j.procs.2021.12.287

1. Introduction

The study of the factors that affect the dropout of students in educational institutions is one of the most frequent topics in the educational field. One of the first studies on this phenomenon was presented by [1], in which a multiple regression model was used to find the level of contribution of a set of social factors concerning the dropout statistics.

In recent years, advanced data mining techniques [2–4] and the use of classification models [5], [6] have become one of the main tools for predicting academic performance [7–9] and student dropout statistics at all educational levels. The most frequent dropout prediction studies involve distance education models and virtual courses in which the rate of dropout is frequently higher [6,10–12].This article aims to show the results obtained through the implementation of a predictive information system (IS) for the desertion of university students. The methodological design consisted of 3 phases: data collection, modeling, and implementation/validation. The resulting IS was implemented in a higher education institution with more than 15,000 students, provides a user interface for monitoring the main risk factors associated with student's dropout, in a timely and personalized way. The article is structured as follows: section 2 shows a description of the methodological phases; Section 3 shows the results and the primary considerations of each of the information gathering, modeling, and implementation phases. Finally, the main conclusions and a set of future research opportunities are listed in section 4.

2. Methods

2.1. Phase 1: Planning and information-gathering.

During this phase, the availability of data was verified in the different departments of the higher education institution. Socio-economic and demographic data, academic performance records, and national standardized test (ICFES) scores were included. The data was collected, consolidated, and complemented with information is obtained from the Colombian Ministry of National Education, and the Colombian Directorate of Social Development).

2.2. *Phase 2: Modeling.*

2.2.1. Variable Selection. Given that the students of each higher education institution have a specific set of characteristics, the dropout patterns may vary according to the institution. A "naive" approach was adopted, aiming to achieve an objective point of view at the time of analyzing the dropout phenomenon. All data were included in the analysis, without any of the variables being previously discarded. This selection process aims to reduce the number of variables that will be used to train the model, prioritizing those with the most significant predictive potential.

2.2.2. Classifier Selection: Instead of predefining a single model that works for most educational institutions, several methods were tested simultaneously, and finally, the method that best suits the reality of the institution was chosen. Thus, the predictive model will be specifically adapted to the dropout patterns observed in the students of the institution. The goal of this step was to determine the classification algorithm with the highest prediction efficiency on the dataset. The algorithms considered during this phase included: AdaBoost algorithms, Bayesian GLM, Decision Trees, Logit Boost algorithms, Random Forest, and Stochastic Gradient Boosting.

2.3. Phase 3: Implementation and Validation of results.

For the implementation of the SATD (for its acronym in Spanish) software, the period of 2018-2 (second semester 2018) was taken as a test period. During this period, a massive upload of the templates and data collected in the different departments involved in the process was carried out. The active students were registered; the platform was launched in the period 2019-1 with the participation of the team in charge of the student follow-up process. The percentage of coverage of the intervention process on the students and its impact on academic permanence was analyzed.

3. Results

3.1 Variable Selection

The data collected in Phase 1 consisted of 44,031 records of 15,805 students, covering four academic periods (2016-1, 2016-2, 2017-1, and 2017-2), grouped in 165 different initial variables. A correlation analysis was carried out using the Kolmogorov Smirnoff (KS) test for each of the variables related to the dropout rates. Subsequently, a multivariate analysis method was used using a Random Forest Algorithm to determine the degree of contribution of various sets of variables in the prediction of dropout rate. Decision trees were generated for the most significant variables. The decision trees provided the ranges or values where the dropout rate is significantly lower or higher than the average (Figure 1). The students were classified in one of the branches (nodes), and in the row depending on the category in which the student is (1: drop out; 0: remain).

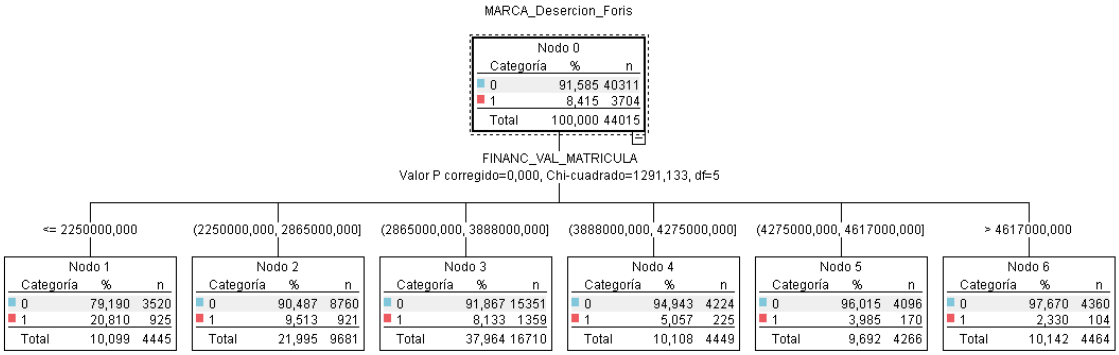

Figure 1. Example of the Decision Trees result (Variable: Tuition Fees in COP).

All the variables were ordered according to their predictive potential. This process requires the use of advanced Data Mining techniques because, for each variable, its conditional dependence on the presence of other variables with respect to dropout must be taken into consideration, and not only its non-conditional dependence. A Random Forest algorithm was used as a regression method on the data (not to be confused with Random Forest in its use as a classification algorithm, which is explained in the classifier selection stage). The process consisted of the following stages:

- Random groups are chosen from the variables present in the dataset, where each variable can be included in more than one tree, and the total of variables per group can vary.
- For each group, a random subset of rows is chosen from the original dataset (student enrollment).
- A decision tree is trained for each group, predicting dropout with the different training data.
- Each variable receives an importance value, which is measured considering the set of decision trees with and without it.
- The variables are ordered according to their importance, and a multivariate ranking is obtained.

3.2. Classifier Selection.

The procedure for the classifier selection started choosing the first variable in the multivariate ranking and training a model for each type of algorithm, predicting dropout based on that single variable. The precision result was measured according to the area under the (Receiver Operating Characteristic Curve) ROC, and the result of each partial model was saved. Then the procedure was repeated with the first two variables in the ranking, then the first three, and so on. The first 30 variables were used at most, to avoid overfitting and to reduce the number of variables that will enter the final result. The different classifiers tested in this stage are described below:

- AdaBoost Algorithm:

AdaBoost is an algorithm to build a "strong" classifier from the weighted sum of a large number of base classifiers, and by default, they are decision trees. AdaBoost classifiers generally have lower predictive quality, but they are easier to obtain. In each iteration, the algorithm learns from the misclassified data, updating the weights that accompany the weak classifiers, and thus improving the predictive quality at each stage [13,14].

- Bayesian GLM:

In the same way as the Logistic Regression, this model focuses on predicting the occurrence of a binary (or dichotomous) phenomenon. However, the parameters of this model are random variables so that probability distributions can be assumed for them before the arrival of data, and the change in their distribution caused by the data considered in the analysis can be known [15].

- Decision Trees:

A tree is made up of a set of nodes, branches, and leaves. Each node represents a variable, which can be divided into branches or leaves. The main decision rule is to separate the data into two groups, as different as possible and to do this, it looks for the variable that best performs this segmentation. All of this process continues until a constraint is met: it can be a fixed number of nodes or fixed depth of the tree. To predict using this method, a "new" student is taken, and how he moves inside the tree is observed, reaching a final decision, which in this case is whether to drop out or not [16].

- LogitBoost Algorithm:

This algorithm uses the same principle as AdaBoost, but the base classifiers used in this process are classifiers created from logistic regressions [17].

- Random Forest

Random Forest algorithm creates a large classifier formed from the development of many small decision trees created randomly. The final prediction is made through the weighted average of the classification of all trees. [18].

- Stochastic Gradient Boosting:

This technique consists of iteratively creating a classifier that, using regressions obtained from sub-samples of the data, learns at each stage of the data that is misclassified, adapting to the observed reality [19].

The area under the ROC curve provides a measure of the performance of the model for the selected variables, equivalent to the probability that the model classifies a real dropout student with a higher probability than a false dropout student (false positive). The two algorithms with the largest area under the ROC curve (close to 80%) was the AdaBoost algorithm and Bayesian GML and provided the best balance between predictive level and model stability. From the initial 165 variables, the first 30 most significant were used to select the classifier. Finally, a set of 25 variables allowed the model to obtain a precise prediction. The probability of dropping out of each student was calculated using the 25 variables. This probability defines the level of priority that will be taken when carrying out activities to monitoring and preventing dropouts by those responsible for the follow-up process. The probability changes according to the alerts and interventions registered by the responsible for the follow-up process during the academic period (Section 3.3.2. and 3.3.3.). At the end of each semester, the impact of the interventions on the percentage of student retention is calculated.

3.3. Implementation.

3.3.1. The information system UI.

The implementation phase began with the uploading of templates to the platform. The responsibility for monitoring each student rests with the faculty welfare coordinators for each program, a full-time professor, and the group of psychologists, advisers, and monitors led by the institution student welfare department. The responsible for the follow-up process can schedule interventions. The calculated dropout probability represents one of the primary decision criteria for those responsible for the monitoring process and is visible information within the application interface.

3.3.2. The process of creating Alerts

Alerts can be classified in three types: Scheduled alerts are those that arise from rules configured in the system (course failure, scholarship holders with a low grades average, withdrawal of subjects). The second type are alerts from the predictive model, which are categorized by a risk factor (academic, personal, or socio-economic) and finally Manual alerts are those created directly by those responsible for the follow-up process.

3.3.3. The process of creating interventions.

Interventions can be created in bulk (in the case of group activities), or individually using the intervention module. These options allow specifying whether the student attended or not and whether the intervention was satisfactory or unfavorable.

3.4. Validation of Results

3.4.1. Interventions Coverage and Dropout rate.

The percentage of coverage of the follow-up process carried out on students increased by 15.2% compared to that observed in 2018 before the implementation of the information system. Approximately 26% of the student population could not be intervened in 2019 (15.6% less than 2018). The reasons can include the presence of outdated contact information in the academic system, and the infrequent use of institutional mail by certain groups of students. Table 1 shows the number of intervened and not intervened students versus their permanence and graduation, or eventual dropout. These data were analyzed with a chi-square test with an alpha value of 0.05, and two hypotheses were established: H_0*: Permanence and dropout are independent of interventions* and H_a*: There is a dependence between permanence/dropout and interventions.* The calculated statistic was 220.04, and the test statistic is 3.84. There is sufficient statistical evidence for not accepting the null hypothesis in favor of the alternate one.

Table 1. Impact of the interventions registered in the student dropout.

Intervened?	Dropped Out	Stayed or Graduate (Enrolled)
Yes	12% (1107)	88% (8026)
No	23% (732)	77% (2450)

Figure 2. Coverage of interventions for dropout risk quintiles.

3.4.2. Dropout Risk Quintiles

The dropout probability calculated by the model allowed the students to be grouped into five risk levels or quintiles ("5" being the highest risk and "1" the lowest risk) (See Figure 2). There is a dependency between the interventions and the permanence of the students, which leads us to conclude that effectively the retention actions that the institution carries out with the students promote their permanence and graduation. Then, we can also validate the effectiveness of the follow-up process of the students through the implemented IS.

4. Conclusions

The implemented IS represents a powerful tool for predicting, monitoring, and managing the risk factors associated with the student dropout factors. Some advantages of the implemented IS include the centralization of the information allowing a comprehensive view of the students, the prioritization in the timely follow-up of students with

a higher risk of dropping out. Also, it allows individual and massive register of the interventions carried out to students. This includes scheduling of future interventions and defining a way to evaluate of the impact of follow-up strategies on student permanence. The limitations of the IS include the need to make more flexible the assignment of a higher number of monitoring managers to generate alerts and improving the customization of reports. An update of the software is proposed that incorporates new functionalities in terms of bulk uploads, increasing data volumes when downloading reports, and improvements in the implementation of the strategy by those responsible, giving priority of intervention to students according to the level of risk using the following order 4, 3, 5, 2 and 1 for greater effectiveness. Among the opportunities for future research is the incorporation of new classification and weighting methods, which allow to improve the reliability of the information system predictions.

References

[1] Spady WG. Dropouts from higher education: Toward an empirical model. Interchange 1971;2:38–62. https://doi.org/10.1007/BF02282469.

[2] Freitas FADS, Vasconcelos FFX, Peixoto SA, Hassan MM, Ali Akber Dewan M, de Albuquerque VHC, et al. IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. Electron 2020;9:1–14. https://doi.org/10.3390/electronics9101613.

[3] Narayanasamy SK, Elçi A. An effective prediction model for online course dropout rate. Int J Distance Educ Technol 2020;18:94–110. https://doi.org/10.4018/IJDET.2020100106.

[4] Patacsil FF. Survival analysis approach for early prediction of student dropout using enrollment student data and ensemble models. Univers J Educ Res 2020;8:4036–47. https://doi.org/10.13189/ujer.2020.080929.

[5] Gopalakrishnan A, Kased R, Yang H, Love MB, Graterol C, Shada A. A multifaceted data mining approach to understanding what factors lead college students to persist and graduate. Proc. Comput. Conf. 2017, vol. 2018- Janua, 2018, p. 372–81. https://doi.org/10.1109/SAI.2017.8252128.

[6] Rolandus Hagedoorn T, Spanakis G. Massive open online courses temporal profiling for dropout prediction. Proc. - Int. Conf. Tools with Artif. Intell. ICTAI, vol. 2017- Novem, 2018, p. 231–8. https://doi.org/10.1109/ICTAI.2017.00045.

[7] Rastrollo-Guerrero JL, Gómez-Pulido JA, Durán-Domínguez A. Analyzing and predicting students' performance by means of machine learning: A review. Appl Sci 2020;10. https://doi.org/10.3390/app10031042.

[8] Hassan H, Anuar S, Ahmad NB. Students' performance prediction model using meta-classifier approach. vol. 1000. 2019. https://doi.org/10.1007/978-3-030-20257-6_19.

[9] Figueroa-Canas J, Sancho-Vinuesa T. Early prediction of dropout and final exam performance in an online statistics course. Rev Iberoam Tecnol Del Aprendiz 2020;15:86–94. https://doi.org/10.1109/RITA.2020.2987727.

[10] Aldowah H, Al-Samarraie H, Alzahrani AI, Alalwan N. Factors affecting student dropout in MOOCs: a cause and effect decision-making model. J Comput High Educ 2020;32:429–54. https://doi.org/10.1007/s12528-019-09241-y.

[11] Skalka J, Drlik M. Automated assessment and microlearning units as predictors of at-risk students and students' outcomes in the introductory programming courses. Appl Sci 2020;10. https://doi.org/10.3390/app10134566.

[12] Viloria A, Padilla JG, Vargas-Mercado C, Hernández-Palma H, Llinas NO, David MA. Integration of data technology for analyzing university dropout. Procedia Comput. Sci., vol. 155, 2019, p. 569–74. https://doi.org/10.1016/j.procs.2019.08.079.

[13] Punlumjeak W, Rugtanom S, Jantarat S, Rachburee N. Improving classification of imbalanced student dataset using ensemble method of voting, bagging, and adaboost with under-sampling technique. Lect. Notes Electr. Eng., vol. 449, 2017, p. 27–34. https://doi.org/10.1007/978-981-10-6451-7_4.

[14] Zeng W, Chin S-C, Zeimet B, Kuang R, Chi C-L. Dropout prediction in home care training. Proc. 10th Int. Conf. Educ. Data Mining, EDM 2017, 2017, p. 442–3.

[15] Feki-Sahnoun W, Njah H, Hamza A, Barraj N, Mahfoudi M, Rebai A, et al. Using general linear model, Bayesian Networks and Naive Bayes classifier for prediction of Karenia selliformis occurrences and blooms. Ecol Inform 2018;43:12–23. https://doi.org/10.1016/j.ecoinf.2017.10.017.

[16] Toivonen T, Jormanainen I. Evolution of decision tree classifiers in open ended educational data mining. ACM Int. Conf. Proceeding Ser., New York, NY, USA: ACM; 2019, p. 290–6. https://doi.org/10.1145/3362789.3362880.

[17] Wang C, Sun J. LogitBoost algorithm considering the cost of misclassification and its application in the classification of mobile user value. Xitong Gongcheng Lilun yu Shijian/System Eng Theory Pract 2019;39:2702–12. https://doi.org/10.12011/1000-6788-2018-0194-11.

[18] Baswardono W, Kurniadi D, Mulyani A, Arifin DM. Comparative analysis of decision tree algorithms: Random forest and C4.5 for airlines customer satisfaction classification. J Phys Conf Ser 2019;1402:066055. https://doi.org/10.1088/1742-6596/1402/6/066055.

[19] Shin Y. Application of Stochastic Gradient Boosting Approach to Early Prediction of Safety Accidents at Construction Site. Adv Civ Eng 2019;2019:1–9. https://doi.org/10.1155/2019/1574297.