Advances in Intelligent Systems and Computing 1366

Álvaro Rocha · Hojjat Adeli ·
Gintautas Dzemyda ·
Fernando Moreira ·
Ana Maria Ramalho Correia *Editors*

Trends and Applications in Information Systems and Technologies

Volume 2

Springer

Modelling academic dropout in computer engineering using artificial neural networks

Diogo M. A. Camelo, João C. C. Santos, Maria P. G. Martins[1,2], and Paulo D. F. Gouveia[1]

[1] Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
[2] CISE – Electromechatronic Systems Research Centre, University of Beira Interior, Calçada Fonte do Lameiro, P – 6201-001 Covilhã, Portugal.

Abstract. School dropout in higher education is an academic, economic, political and social problem, which has a great impact and is difficult to resolve. In order to mitigate this problem, this paper proposes a predictive model of classification, based on artificial neural networks, which allows the prediction, at the end of the first school year, of the propensity that the computer engineering students of a polytechnic institute in the interior of the country have for dropout. A differentiating aspect of this study is that it considers the classifications obtained in the course units of the first academic year as potential predictors of dropout. A new approach in the process of selecting the factors that foreshadow the dropout allowed isolating 12 explanatory variables, which guaranteed a good predictive capacity of the model (AUC=78.5%). These variables reveal fundamental aspects for the adoption of management strategies that may be more assertive in the combat to academic dropout.

Keywords: Educational data mining, Artificial neural network, Academic dropout, Predictive Model

1 Introduction

Due to the importance of education for the progress of society as well as for citizens' well-being and prosperity, effective measures are required to combat high school dropout in the current context of higher education institutions (HEI). In the majority of studies on the portuguese HEI, about the factors of success and academic failure, an exploratory and descriptive record is more privileged than an explanatory and propositional record ([1]) which indicates a poor prognosis for the design of effective early interventions. In order to assist the institutional decision-makers in this process, a predictive classification model is presented that allows the prediction, at the end of the first school year, of the dropout tendency of a student of computer engineering (CE) of the Polytechnic Institute of Bragança (IPB). The main explanatory factors for dropping out were identified with the help of data mining techniques and artificial neural networks (ANNs), applied to records of 653 CE students of the IPB, characterized by 22 factors of students' academic and demographic dimensions and, in particular, based on intermediate (un)success patterns of their academic career. By presenting a model with a considerable success rate (AUC=78.5%), when applied to real data, the knowledge obtained may prove crucial to improve decision-making to combat student dropout.

After this introduction, the present paper is composed of the following sections Sect. 2 – outline of related studies; Sect. 3 – presentation of the methodology and of the data model developed; Sect. 4 – presentation and discussion of results of the prediction model proposed; Sect. 5 – final conclusions and perspectives of future work.

2 Educational Data Mining

The use of data mining techniques in modelling performance and academic dropout is a promising area of research called Educational Data Mining (EDM). Major reviews of the state of the art [2–5] prove the importance and usefulness of EDM, as a tool for analysis and management support. According [6] the main goal of EDM is to generate useful knowledge which may ground and sustain decision-making targeted at improving student communities's learning as well as educational institution's efficiency. In fact, most EDM studies are subject to performance prediction and academic dropout, which identify the factors that can influence them and constitute fundamental aspects in the definition of management strategies focused on promoting success and preventing school dropout [5]. Whithin this context, the main aim of the studies [7–9] was to conduct a review to identify the most commonly studied factors that affect the students' performance, as well as, the most common data mining techniques applied to identify these factors. In the study by [7] was pointed out that the most common factors are grouped under four main categories, namely, students' previous grades and class performance, e-Learning activity, demographics, and social information. Additionally, Shihari [8] showed that the cumulative grade point average, and internal assessments (marks after entering higher education such exams, assignments, etc.) are the most frequent attributes used for predicting the student's performance. Furthermore, other important attributes were also identified, including student's demographic and external assessments (pre-university achievement classifications), extra-curricular activities, high school background, and social interaction network. Also, from the analysis of 10 EDM studies with an emphasis on foreseeing academic dropout,the author [9] concluded that in most cases, the average access to higher education, the level of education, and the parents' profession and poor methodology are the main factors that affect academic dropout. Among the DM techniques most used these studies concludes that the Decisin Tree, Artificial Neural Networks (ANN), Naive Bayes (NB) were, in descending order, the DM techniques most often used for forecasting purposes in EDM ([7, 8]).

Regarding school dropout, the scope of the present study, the authors of the study [10], through the development of models that integrated the Random

Forest (RF), support vector machine(SVM) and ANN techniques, concluded that there are essentially 2 factors related to students' curricular context which mostly account for the academic dropout of the undergraduates attending the 50 degree courses taught in the institution used as a case study. The models presented, whose predictive accuracy was 76% when applied to new data, were developed with records of 3344 students, characterized by more than 4 dozens of potential predictors of dropout concerning students' academic and demographic dimension, and their socio-economic status.

In the study [11] considering demographic, socio-economic and learning performance information of 972 students at Riddle Aeronautical University, processed by learning algorithms, Naive Bayes, k-Nearest neighbors, RF, ANN, decision trees and logistic regression, concluded that the grades average, the classifications in the entrance exams, the average grade of the first year of college and the financial contribution provided by the family are, in descending order, the factors that best explain the student dropout.

Classification rules and the algorithms, Naives Bayes, support vector machines, instance based Lazy Learning and Jrip were used in the study by [12], which results showed, with success rates above 80%, that it was possible to identify the propensity to drop out of high school students in Mexico. In the development and validation of the proposed model, information from 419 students was used in a very comprehensive way in academic and socioeconomic dimensions.

3 Data and Methodology

3.1 Data model

The dataset that supports the creation of the model capable of predicting the academic dropout of the students of CE at Institute Politecnic of Bragança was obtained from the Information System of that same educational institution and contains records of 635 students, enrolled in CE between 2006 and 2019.

In the data selection and pre-processing phases, which preceded the applications of ANNs, there was a need to clean and transform the data. In particular, removing the enrollment of students from the pre-Bologna study plans, of those who have changed their course or educational institution, of those who are still enrolled in this school year (still of undefined outcome) and also, for each student, data was deleted from all enrollments subsequent to the first (since the prediction is only made with information available until the end of the 1st year). The target variable of prediction is of boolean type (1 or 0), whose value can take one of the following meanings: 'dropout', if the student did not complete the degree, or 'not dropout' if they concluded the degree. All students who did not have a valid registration in the year 2019/2020 and who, simultaneously, had not yet completed their course, had their enrolment classified as 'dropout'. As the main operations of variable transformation, it was necessary to normalize the predictors of the numerical type, using, for this purpose, the Min-Max scaling approach, through which all the values of the numerical variables became included between 0 and 1. There was also the need to calculate new variables, such as, for example, the predictive variable 'n_ects_done', which was obtained from the curricular units in which the student was approved.

After the cleaning of the data and the performing of other pre-processing tasks, each of the IE students was characterized by a total of 22 potential explanatory variables of dropout, categorized into 3 main groups: demographic (D), curricular (C) and of matriculation (M). The variables in question are represented in Table 1.

Table 1. Predictive variables considered in the abandonment prevision.

attribute	type	category	meaning
gender	nominal	D	student's gender
nationality	nominal	D	student's nationality
cod_district	nominal	D	code of the district where the student lives
cod_district_n	nominal	D	code of the district where the student was born
ALGA	discrete	C	grade in Linear Algebra and Analytical Geometry
CI	discrete	C	grade in Calculus I
F	discrete	C	grade in Physics
PI	discrete	C	grade in Programming I
SD	discrete	C	grade in Digital Systems
AC	discrete	C	grade in Computer Architecture
CII	discrete	C	grade in Calculus II
MD	discrete	C	grade in Discrete Mathematics
PII	discrete	C	grade in Programming II
age	discrete	D	age at first registration
phase	ordinal	M	phase of the entry into the degree
cod_type_entry	nominal	M	code of the type of entry of the student at the degree
scholarship	logical	C	where the student had a scholarship
associative_leader	logical	C	where the student was associative leader
cod_freq_type	nominal	C	code of the type of frequency of the student
cod_status	nominal	C	type of the status of the student
n_ects_done_auto	discrete	M	number of credits done automatically on degree entry
n_ects_done	discrete	C	number of credits completed done

3.2 Methodology

In the development of the model proposed, an approach using Artificial Neuronal Networks was chosen because of the ability and proven effectiveness demonstrated in some studies in the area of EDM (e.g.[3, 4, 7]).

All of the computational calculations inherent to the study are performed in R, in the Rstudio environment and specific packages for applying ANNs are used as an extension to R itself. All the pre-processing that preceded the application of the ANNs in the development of the new predicting model was carried out in Mysql.

In the assumption of the established objectives, a random partitioning of the dataset is performed in 3 subsets: training, validation and testing. Such partitioning is characterized in Table 2.

Table 2. Dimensions of the subsets used for training, validation and testing.

total	train	validation	test
653	391	130	132

With the training and validation subsets, and the processes of learning and refining, respectively, the ANNs is performed. Once trained and properly refined, from the initial set of 22 potential predictive variables, the identification of the most explanatory factors of academic dropout is conducted, through the selection of those that show greater impact to the intended prediction (feature selection). For that purpose, a progressive selection method (forward search) is adopted, in which the variables are selected one by one, in an iterative process, always joining to the ones previously selected the remaining ones that lead to a greater increment. Of course, the process ends when this increment becomes null or negative.

After the set of variables most explanatory of dropout has been identified, the influence of each of them in the target variable of prediction is measured by means of a simple technique of sensitivity analysis, which attempts to define the importance of each variable by measuring the loss of accuracy resulting from its non-inclusion in the model.

The subset of data left for testing is used only in the final assessment of the model in order to measure its generalization capacity. As a predictive evaluation metric for the different ANN configurations, the Area Under ROC Curve (AUC) is used, in which the ROC curves (Receiver Operating Characteristic) form a graph that illustrates the performance of a binary classification model, based on the specificity and sensitivity of the same classifier model.

4 Implementation and results

For an easier understanding of the results presented in the next sections, a scheme is shown in Fig. 1 intended at illustrating the 2 predictive classification models developed and, in particular, the different subsets of predictors that support them.

4.1 Training and refinement of ANNs with all the independent variables

In order to obtain a first perception of the predictive capacity of the model, the first step consisted of training the model with the complete set of independent variables available. For easier referencing, the predictive model that integrates

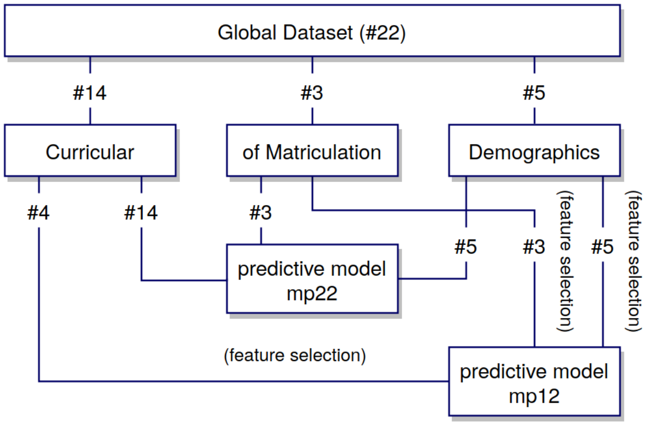

Fig. 1. Illustrative diagram of the predictive model.

all these variables will be named *mp22*. As recommended in the literature, and with the objective of maximizing the predictive capacity, the model was refined, through the evaluation of its performance for different values of its most important hyperparameters. In the specific case of the present study, the hyperparameters used were the size (number of neurons of the hidden layer of ANNs) and Decay (learning rate decay). The 10 best values obtained in this refinement process can be found in Table 3. The prediction model was then supported by an ANN of 6 neurons (in its hidden layer) and with a decay rate of $10^{-\frac{1}{3}}$.

Table 3. The 10 best refinement results.

size	6	4	8	2	3	3	10	20	5	15
decay	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{2}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$	$10^{-\frac{1}{3}}$
AUC	0.847	0.846	0.846	0.846	0.844	0.844	0.844	0.843	0.843	0.843

4.2 Selection of the main explanatory factors of dropout

After completing the training and refinement phases of the predictive model, an adjustment of the set of variables that support the model was made with the same training and validation data, by selecting, through the forward search method, those that had proven to be the most accountable ones for dropout. The result obtained by the iterative process of variable selection is reported in Table 4.

From the table, it is possible to conclude that the number of variables most explanatory of academic dropout in CE is 12, from which 4 belong to the students' curricular dimension, 3 represent matriculation data and 5 are of demographic nature. This new model will henceforth be called *mp12*.

Table 4. Variables selected by applying the forward search method.

order	atribute	type	category
1	phase	ordinal	M
2	AC	discrete	C
3	age	discrete	D
4	n_ects_done	discrete	C
5	cod_type_entry	nominal	M
6	F	discrete	C
7	cod_district_n	nominal	D
8	cod_district	nominal	D
9	n_ects_done_auto	discrete	M
10	gender	nominal	D
11	nationality	nominal	D
12	MD	discrete	C

4.3 Evaluation of the generalization capacity of the model found

After the training, refinement and feature selection processes were completed, the new *mp12* model was evaluated with the test subset in order to verify its true generalization capacity. The results of the respective previsions are shown in Table 5, which also includes, for comparison purposes, the results previously obtained with the validation data.

Table 5. Comparison of the models performance.

model	AUC (validation)	AUC (test)
mp22	84.7%	76.7%
mp12	85.9%	78.5%

The comparison between the AUC obtained with the test data and what had been obtained with the validation data allows seeing that, as expected, the AUC values decrease – 8% in the model that uses all the variables available (*mp22*) and 7.4% in the model where the most explanatory variables were selected (*mp12*). These results reveal that the reduction of variables has greatly increased the generalization capacity of the model. In addition to having fully achieved what is the supreme purpose of this type of study (maximizing the capacity for generalization), the reduction of variables also translates a set of other advantages when the model is applied in real contexts, namely, by allowing to identify the main factors that explain the target predictive variables, helping eliminate redundant variables, decreasing the computational complexity of the model and facilitating its application in real contexts.

In order to have another perspective on the behavior of each dropout prediction model, the respective ROC curves are overlapped in Fig. 2. Although the differences in performance between the two models are not easily perceptible in the figure, it is indeed possible to notice that the curve that is closest to the vertex characterizing the condition of optimality (specificity = 100% and sensitivity = 100%) is the model with the fewest variables (*mp12*).

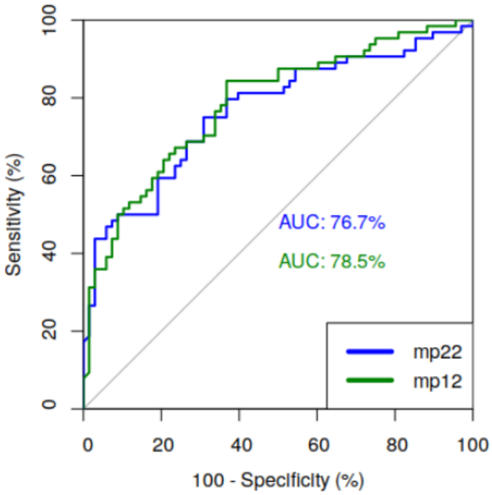

Fig. 2. Final ROC curves of the models.

4.4 Relative importance of the main explanatory factors

After finding the set of variables which are most explanatory of dropout, and in order to distinguish the influence of each of them, an order of relevance was established among them, according to their importance for the prediction. The Fig. 3 shows the values of the relative importances in question.

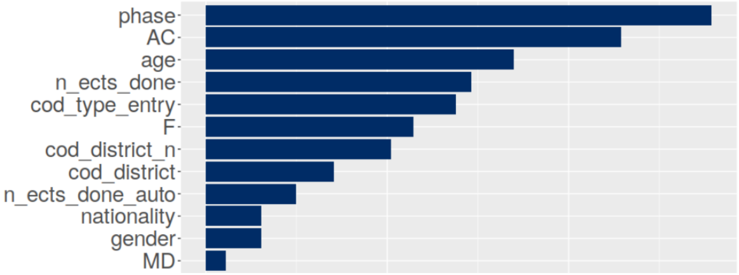

Fig. 3. Importance of the predictive variables.

It is relevant to point out the great influence that the 'entry phase', the moment in which the student entered the course, has on the explanation of academic dropout. Other variables that also reveal great explanatory power are the classification in the course unit of 'computer architecture' and the 'student's age' at the time of entering the course. It would be expected, however, that the variables 'number of ECTS completed at the end of the first school year' and 'number of ECTS automatically credited upon entry' would have shown greater explanatory power regarding dropout. The reason why this did not occur is most likely due to the high degree of correlation between these variables and other potential predictors of dropout, such as those related to the classifications obtained in certain course units.

5 Conclusions and future work

With the application of data mining techniques and, in particular, of artificial neural networks, a model of classification was developed, which allows predicting, at an early stage of the academic career, whether an IPB computer engineering student will be an academic dropout or a graduate. From a set of 22 predictive factors of dropout, the process of knowledge discovery in a database led to the conclusion that 12 factors are the most explanatory of dropout in CE.

Subsequently, a sensitivity analysis technique was applied to the model with the 12 variables in order to calculate the relative importance of each one of them, after which it was found that the 'phase of entry into the course', 'the classification in the course unit of computer architecture' and the 'student's age at the moment of entering the course' are the 3 variables with greater influence on the dropout of CE students.

Since the classification obtained in the course unit of computer architecture is a strong predictor of dropout, some possible future work to be considered is the development of a predictive regression model that allows estimating students' grades in this course unit. With such knowledge, institutional decision-makers could define ways to promote the educational success of students for whom a lower performance may have been estimated by the model.

In order to complement this research, the cross-validation technique could be used, for instance, in the partitions of training, validation and testing. The results obtained in this study may support the definition of effective measures to combat academic dropout in the Computer engineering degree course, where dropout rates are worrisome, since they reach about 50% in the dataset analyzed. For instance, setting up study support groups for the course unit of computer architecture might be one of the guidelines to suggest. Since EDM literature demonstrates that decision trees and artificial neural networks are the most widely used learning techniques in student modelling, it will be possible to explore the relevance of using other techniques in future work, that also often display very competitive results, such as committee-based ones, also called ensembles methods or mixture of experts.

Acknowledgments This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Projects UIDB/04131/2020 and UIDP/04131/2020.

References

1. Costa A., Lopes J. and Caetano A.: Percursos de estudantes no ensino superior. Fatores e processos de sucesso e insucesso. Lisboa: Mundos Sociais. pp. (2014)
2. Bakhshinategh,B., Zaiane, O. R., Elatia, S., and Ipperciel, D.: Educational data mining applications and tasks: A survey of the last 10 years. Education and Information Technologies, 23(1), pp. 537-553. (2018).
3. Peña-Ayala, A.: Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works. Expert systems with applications. 41(4), 1432–1462, 2014.
4. Romero, C., Ventura, S.: Educational Data Mining: a Review of the State of the Art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 40(6), 601–618 (2010).
5. Martins, M.P.G., Migueis, V.L., Fonseca, D.S.B.:Data mining educacional: uma revisão da literatura. 13th Iberian Conference on Information Systems and Technologies (CISTI). Institute of Electrical and Electronics Engineers. (2018)
6. Martins, M.P.G., Migueis, V.L., Fonseca, D.S.B., Alves, Albano.: A data mining approach for predicting academic success–A case study. International Conference on Information Technology & Systems. Springer. Vol. 918, pp. 45–56. (2019)
7. Saa, A. A., Al-Emran, Mostafa and Shaalan, Khaled: Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. Journal of Technology, Knowledge and Learning. Springer. Vol. 4, pp. 567–598 (2019)
8. Shahiri, A. M., Husain, Wahidah and others:A review on predicting student's performance using data mining techniques.Procedia Computer Science. Elsevier Vol. 72, pp. 414–422. (2015)
9. Kumar, A. D., Selvam, R. P., Kumar, K. S.:Review on prediction algorithms in educational data mining.International Journal of Pure and Applied Mathematics. Vol. 118, pp. 531–537. (2018)
10. Martins, M.P.G., Migueis, V.L., Fonseca, D.S.B., Gouveia, P.D.F.: Previsão do abandono académico numa instituição de ensino superior com recurso a data mining. Revista Ibérica de Sistemas e Tecnologias de Informação. Associação Ibérica de Sistemas e Tecnologias de Informacao. Vol. E28, pp. 188–203. (2020)
11. Lehr S., Liu H., Kinglesmith S., Konyha A., Robaszewska N., and Medinilla J.: Use educational data mining to predict undergraduate retention. In Advanced Learning Technologies (ICALT), IEEE 16th International Conference on Technologies, Vol. 61. pp. 428–430. IEEE, (2016)
12. Márquez-Vera C., Cano A., Romero C., Noaman A., Fardoun H and Ventura S.: Early dropout prediction using data mining: a case study with high school students. Expert Systems, 33(1) pp. 107–124, (2016)