Using Academic Performance to Predict College Students Dropout: a case study[1]

António Carlos Corte-Real de Sousa[2]
Carlos Alberto Bragança de Oliveira[2]
José Luís Cabral Moura Borges[2]

Abstract

Student dropout is a complex problem that affects most post-secondary undergraduate programs, all over the world. The Industrial Engineering program of the ISVOUGA Institute, located in Sta. Maria da Feira, Portugal, is no exception. This research used a dataset containing students' general information and the students' marks for the already assessed courses. From this dataset, 17 potential predictors have been selected: five intrinsic predictors (gender, marital status, professional status, full/part time student, and age) and 12 extrinsic ones (the marks in all the 12 courses taught during the first two semesters of the program). The main goal of this research was to predict the likelihood of a student to dropout, based on the referred predictors. A binary logistic regression was used to classify students as having a high or low probability not to re enroll the program. To validate the appropriateness of the used methodology, the accuracy of the logistic model was compared, by means of a 5-fold cross-validation, to the accuracy of three classification methods commonly used in Data Mining: One R, K Nearest Neighbors, and Naive Bayes. Four variables were significant to the logistic model (the marks in Materials Science, Electricity, Calculus 1, and Chemistry). The two most influential predictors for student dropout are failing to pass in the less challenging courses of Materials Science and Electricity. Contrary to what we would think prior to this research, we found that failing in more challenging courses such as Physics or Statistics does not have a significant influence on student dropout.

Keywords

Student dropout – Retention – Logistic regression – Data mining.

1 - Acknowledgments: We would like to thank Prof. Teresa Leão, head of the *Instituto Superior de entre o Douro e Vouga* (ISVOUGA), for giving us access to the Institute databases and for her insights and discussion on student dropout. This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project "POCI-01-0145-FEDER-006961", and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia as part of project "UID/EEA/50014/2013".

2- Universidade do Porto, Porto, Portugal. Contact: a.sousa@doc.isvouga.pt; ORCID: http://orcid.org/0000-0002-6493-6161; braganca@fe.up.pt ORCID: http://orcid.org/orcid.org/0000-0002-9505-8170; jlborges@fe.up.pt; ORCID: http://orcid.org/0000-0001-9946-5614.

DOI: http://dx.doi.org/10.1590/S1678-4634201844180590

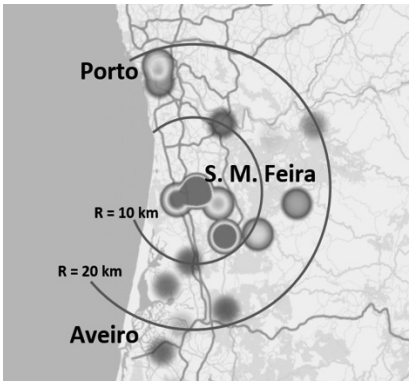
Introduction

Education is a challenging subject for most countries. In 2016, Portugal was still below the EU22 level of educational attainment for the 25-34 year old range, having a rate of attainment for post-secondary education of 35%, while the corresponding EU22 average rate was 41%. Regarding a wider range of ages (25-64 year-old), 24% of Portugal's adult population had attained post-secondary education in 2016, which is below the OECD average of 37% (OECD, 2017).

The ISVOUGA[3] Institute is a private post-secondary undergraduate school, having 350 students in 2015. It is located in Santa Maria da Feira, which is a small village with a population around 12,500 (INE, 2011). Santa Maria da Feira is located in Portugal, 25 km South from Porto and 30 km North from Aveiro. Porto is the 2nd largest Portuguese city with a population of 250,000 (INE, 2011), having 40 post-secondary public and private schools, one of which is the public University of Porto: 30,000 students with a university rank in range 301-400 of Shanghai Ranking Consultancy (ARWU, 2016). Aveiro is a medium-sized village with a population of 65,000 (INE, 2016), having five post-secondary public and private schools, one of which is the public University of Aveiro with 12,000 students and a university rank in range 401-500 of Shanghai Ranking Consultancy (ARWU, 2016). The region in which ISVOUGA Institute is located belongs to a wider region (*Entre o Douro e Vouga*).

Students of ISVOUGA come from villages and cities located within a radius of 25 km from Santa Maria da Feira and out of these, over 90% live at less than 10 km from ISVOUGA (see the origin of the students in Figure 1). In average, between 2010 and 2014, in the *Entre o Douro e Vouga* region, only 2.3% of the resident population with an age in the range 18-22 was attending post-secondary undergraduate programs, while the corresponding national average was 31.8% (INE, 2016). Such small value means that most youngsters of this region drop out of school at the end of secondary education without attending any type of post-secondary courses. In 2015, the average age of the ISVOUGA's freshmen was 25.1 years. This large average value for the freshmen's age suggests that most students had previously left school earlier in their lives, returning to school several years later.

Figure 1- Origin of the students of ISVOUGA Institute



Source: Authors.

3- ISVOUGA is the acronym for Instituto *Superior de Entre o Douro e Vouga*.

Attending a post-secondary program is nowadays a common expectation for people with diverse cultural and social backgrounds (MÍNGUEZ; SAN JULIÁN, 2013; PÁRAMO FERNÁNDEZ et al., 2017). However, it may be a challenging task for most students given that it involves a variety of both intrinsic (students' related) and extrinsic factors (related to the school, social background, and economic background). Student dropout is a complex problem that affects most post-secondary undergraduate programs.

The characteristics of the *Entre o Douro e Vouga* region create a challenging pressure on the ISVOUGA management board. On the one hand, they are urged to identify and understand the external issues that may discourage freshmen to register into the Institute; on the other hand, the internal issues that may lead to the dropout of current students. In public institutions, the high dropout rates result in waste of taxpayers' money and in a population with lower education having, consequently, fewer employment opportunities for positions requiring high qualifications (PAURA; ARHIPOVA, 2014; LITALIEN; GUAY, 2015). In a private institution like ISVOUGA, which is not supported by public funds, high dropout rates do not result in waste of taxpayers' money but represent a waste of students' personal resources (time and money). To avoid the voluntary dropout of students (i.e. when students decide not to re-enroll), the ISVOUGA Institute is particularly careful in defining policies that may help to monitor and support the school's freshmen.

The explanation and prediction of students' academic performance have been widely researched, mainly since the 1980s. Several studies focus on theoretical concepts related to these subjects. For example, PASCARELLA; TERENZINI (1980), ASTIN (1984, 1993), KUH (2003) and others have studied the relationship between academic, social, personal and emotional characteristics to student adjustment and student retention. These researches are useful for Institutional assessment and for a social understanding of the subject but are not easily transposable to definable correcting actions that enhance integration or engagement. Tinto (1982) proposed several types of reasons that influence the likelihood of permanent withdrawal from higher education: student finances; gender; academic and social integration; the interests, skills, values, and commitments to the goals of higher education and to the specific institution; students' skills (academic, social, or otherwise) and/or intellectual capacities; students' interest, commitment and motivation to finish a program.

According to Tinto's model, a higher degree of integration is directly related to a higher commitment to the educational institute and to the goal of study completion. Stratton, O'Toole, and Wetzel (2007) concluded that factors such as the enrollment in full-time or part-time are not the most important factors in the decision to drop out.

After the identification of the students who are at risk of dropping out Neild, Balfanz, and Herzog (2007) as well as Litalien and Guay (2015) suggest several strategies that can help to keep these students on the path to graduation: motivational resources, psychological support provided by advisors, and the influence from faculty or other students in order to reduce retention rates and the dropout intentions. Additionally, peer mentoring by successful students can model studying habits for the freshmen, resulting in an increase in approval rates, social integration and engagement with the

University community (MORALES; AMBROSE-ROMAN; PEREZ-MALDONADO, 2016). Using a not so different perspective from peer mentoring, some other studies identify supplemental instruction as a complementary method to help students improve their performance in 'difficult' courses (MALM; BRYNGFORS; MÖRNER, 2012; MARTÍNEZ-LÓPEZ et al., 2014; MALM; BRYNGFORS; MÖRNER, 2015). Supplemental instruction is a program consisting on collaborative activities under the guidance of a senior student. Some authors also suggest improvements in introductory programs for new students and changes in the evaluation methods (HOVDHAUGEN, 2011).

The ISVOUGA's Industrial Engineering Program takes six semesters in its first cycle program (Bachelor's level) according to the Bologna Working Group on Qualifications Frameworks (2005). It is based on Mechanical Engineering and corresponds to 180 ECTS credits. Students can apply to this program if they have successfully completed their upper secondary (high school) education. However, students older than 23 can apply to this program without having completed their upper secondary. In this latter case, students must submit their curriculum to the scientific board of the Institute, to take an assessment exam and attend an interview with an evaluation committee. The Industrial Engineering program is one of the five bachelor programs offered, at present time, by the ISVOUGA Institute. This program provides the fundamental knowledge and skills required by an Industrial Engineer in areas such as Mathematics, Physics, Design and Manufacturing, Materials Science, Solids and Fluid Mechanics, Statistics and Data Analysis. In order to obtain an objective assessment of the intrinsic level of difficulty of each course for semesters 1 and 2 of the program, a score for the difficulty level of each course was computed. This score has two components: the historical standardized average grades for each course with the standardized number of years until "Pass" for that course. The final score is obtained by the weighted average of the two components considering a weight of 40% for the historical standardized average grades and a weight of 60% for standardized number of years until "Pass". The structure of the first year of the Industrial Engineering program and the corresponding scores are presented in Table 1.

Within the 2009-2013 time window, in the Industrial Engineering program, the average global grade for the students who have graduated was 12.98 (in a 20-point scale) and it took them, in average, 7.38 semesters to successfully complete all the courses taught along the 6 semesters. In the time period 2007-2013, an average of 37 freshmen registered yearly in the program. The most challenging and historically considered as "difficult" courses of the first curricular year (semesters 1 and 2) are: Calculus 1, Physics and Chemistry. The average dropout rate was marginally over 30% of the freshmen who have registered the program and most dropouts (66%) happened at the end of the 2nd semester.

Table 1- Year 1 courses for semesters 1 and 2. A global level of difficulty for each course is presented in column 5

Course	Semester	Average grade [0-20]	Average number of trials until "Pass"	Level of difficulty of the course * 1 - Easiest, 4 - Most difficult
Geometry and Linear Algebra	1	12.3	1.7	1
Materials Science	1	12.1	2.5	2
Physics	1	11.2	3.8	4
Spreadsheets (MS Excel)	1	13.4	3.0	3
Technical Drawing 1	1	13.8	1.1	1
VB Programing	1	13.2	2.7	3
Calculus 1	2	11.7	4.8	4
Chemistry	2	11.2	4.5	4
Electricity	2	12.9	1.6	2
Statistics	2	11.9	3.2	3
Operations Research	2	12.1	2.1	2
Technical Drawing 2	2	12.7	1.4	1

*Based on the rank of the weighted sum of the standardized average grades (weight 40%) with the standardized number of years until "Pass" (weight 60%).
Source: Authors.

The aim of this research was the analysis of the factors which, at the school level (academic performance), increase the likelihood of a permanent voluntary withdrawal from the Industrial Engineering program.

In this research a student was considered a dropout when he/she has not ended graduation, has obtained at least six final grades in any of the courses of the program and also complies with one of the two following conditions:

- he/she has not renewed or has annulled their registration or;
- he/she did not attend any evaluation for any course for a time period longer than one year.

We focused the analysis of student dropout in this program because it has an overall stable structure, only with small changes since its beginning in 1996, and because the authors of this research are Industrial Engineers, which gives them an important insight on the characteristics of most of the courses and of the particularities of the Industrial Engineering program.

As a case study, this research stayed focused on the students' academic performance in the two first semesters of the program (the categorical classifications – pass / fail to pass – in the 12 courses taught during the first two semesters of the program) and in the student's intrinsic characteristics (gender, marital status, professional status, full time/part time student, and age at the beginning of the program). This research was not intended to analyze social, economic or psychologic factors that may influence the student dropout (e.g., student's socio-demographic background, motivations for studying, social and academic integration at school, living conditions). These factors have been shown to be related to high rates of dropout (JORDAN; LARA; MCPARTLAND, 1994; PIERRAKEAS et al., 2004; PAURA; ARHIPOVA, 2014).

Out of the 17 potential predictors selected from the database with students' information only the extrinsic factors (the students' marks), if included in the model, would be transposable to practical actions that might reduce the likelihood of voluntary withdrawal. Intrinsic characteristics cannot be changed but have been included as potential predictors in order to assess whether they have influence on the results. The logistic model did not select any of the intrinsic characteristics/predictors as being significant to students' dropout.

The outcome of this research was the identification of a list of first-year courses that mainly accounts for the students' dropout from the Industrial Engineering program and a methodology that enables a probabilistic anticipation of the students who have a strong likelihood of dropping out in the future, based on the students' marks on those courses. These findings provided the management board of ISVOUGA with an invaluable insight to approach the dropout issue.

Data and methods

Data

As a case study, the target population of this research is the set of students attending the Industrial Engineering program of the ISVOUGA Institute. In order to create a prediction model for the first-year students' dropout we had access to the data in the Institute academic database. The records of the students that attended this program from 2007 to 2013 have been used as sample data (192 students' records – a set of 154 records used for training and a set of 38 records used for validation), after having being cleaned[4].

The Institute database contains students' personal data (selected course, name, gender, marital status, professional status, full time / part time student, age at the beginning of the program), the dates and the final marks for the courses for which students had already been assessed by October 2013. First, we anonymized the data, removing all personal information not relevant for the aim of this work. Data was then pre-processed by cleaning erroneous / incomplete records, and only the records where there were at least six final grades in any of the courses of the entire sixsemester program were kept. As a result of the pre-processing only 192 from the initial 310 records were kept.

4- Process of detecting, diagnosing, and editing faulty data.

From the 192 records, 87 (45.3%) were classified as Y=1 (student who dropped out of the program) and 105 (54.7%) were classified as Y=0 (student who did not drop out of the program). Seventeen variables have been considered as potential predictors/regressors (see Table 2). The first five predictors (x_1 to x_5) represent intrinsic factors (student-related) and the remaining twelve (x_6 to x_{17}) represent extrinsic factors (institutional-related).

Table 2 - Variables that have been considered as potential predictors of the dependent variable

	Variable	Description	Type
	X_1	Gender	Categorical
	X_2	Marital status	Categorical
	X_3	Professional status	Categorical
	X_4	Full or part-time student	Categorical
	X_5	Age at the beginning of the program	Numerical
1st year marks in a 20-point scale	X_6	Geometry and Linear Algebra	Numerical
	X_7	Materials Science	Numerical
	X_8	Physics	Numerical
	X_9	Spreadsheets (MS Excel)	Numerical
	X_{10}	Technical Drawing 1	Numerical
	X_{11}	VB Programing	Numerical
	X_{12}	Calculus 1	Numerical
	X_{13}	Chemistry	Numerical
	X_{14}	Electricity	Numerical
	X_{15}	Statistics	Numerical
	X_{16}	Operations Research	Numerical
	X_{17}	Technical Drawing 2	Numerical

Source: Authors.

Multicollinearity between variables was analyzed using the Variance Inflation Factor (VIF). One rule of thumb indicates a VIF value of 5 to be used as a threshold, indicating that multicollinearity may exist (MENARD, 2001). All the 13 numerical variables have a VIF smaller than 3.1 indicating a moderate level of multicollinearity between variables, enabling all the variables to be used as potential predictors.

Methods

Since the output *Y* is categorical having two possible outcomes (1 - student who dropped out of the program, 0 - student who did not drop out of the program), the approximation of P(Y|X) is a classification problem. We performed a binary logistic regression to classify students as 0 or 1. In this context, a binary logistic regression increases the probability that a student belongs to class 1 conditional on the values of several regressors *X*, according to equation (1).

Equation (1)

$$p(class = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta X)}}$$

The logistic regression uses the maximum likelihood method to estimate the coefficient values for $\beta_1, \beta_2, ..., \beta_n$. When the regressor X_j with a coefficient β_j increases by one unit, controlling for the other variables, the odds, $p/(1-p)$, increase by a multiplicative amount of $e^{\beta j}$, where p is the probability associated to class 1.

The variables given in Table 2 are the regressors initially used in the model. After running the logistic model using the variables in Table 2 we repeat the logistic regression using a set of new defined variables named x'_6 to x'_{17} instead of x_6 to x_{17} (courses' marks in a 20-point scale). This new set of variables (x'_6 to x'_{17}) are the variables x_6 to x_{17} transformed from a numerical to a categorical scale, using the following transformation: if $x_i \geq 10$ then $x'_I = 0$ (the student passed the course), if $x_i < 10$ then $x'_I = 1$ (the student failed in the course). The model using categorical variables produced an overall level of accuracy in classifying the students similar to the model that used numerical variables. Since the interpretation of the results of the regression model using categorical variables to represent marks of the 1st year courses was easier than the interpretation of the model using actual numerical marks, we proceed this research referring to the former. The sample size, N=192 students, was larger than the recommended size of at least 100 observations for logistic regression models (LONG, 1997) providing us with a high level of confidence in the interpretation of the regression coefficients.

Additionally, in order to validate the appropriateness of the used methodology to this research, three classification methods commonly used in Data Mining (HAND, 2007) have been employed to compare their overall level of accuracy with the level of accuracy obtained by using binary logistic regression: *i*) One R, *ii*) KNN - K Nearest Neighbors and *iii*) Naive Bayes. One R is a basic methodology that induces classification rules based on the value of a single predictor, K nearest neighbors (COVER; HART, 1967) is an algorithm that classifies data based on a distance (the Value Difference Metric (WILSON; MARTINEZ, 1997) was used in this research) and the Naive Bayes classifier is based on applying Bayes' Theorem with strong independence assumptions which assume the conditional independence of all effect variables (RUSSEL; NORVIG, 1995). Cross-validation was used in order to compare the relative performance of the four methods (REFAEILZADEH; TANG; LIU, 2009). In the cross-validation we have used five folds since the size of the sample was too small to use a larger number of folds and we kept the proportions of Y=1 and Y=0 in each fold balanced with the whole sample correspondent values maintaining, approximately, the proportions of 45.3% of Y=1 (student who dropped out of the program) and 54.7% of Y=0 (student who did not drop out of the program). In the five cross-validation folds the 192 data sample was partitioned into two sets of data; a set of 154 records used for training and a set of 38 records used for validation. The final logistic model was built using the whole 192 data sample and the goodness-of-fit and accuracy statistics were calculated using this model.

Results

In this section we present the results of the logistic regression model. As overall goodness-of-fit statistics we present:

i) the logarithm of the likelihood function associated with the intercept-only model;
ii) the logarithm of the likelihood function associated with the full model (the model that includes the independent predictors as well as the intercept);
iii) the likelihood ratio statistic; and
iv) McFadden's $R^{2.}$

A forward stepwise regression method was used to select the predictors. For each predictor the p-value ($Pr>\chi^2$) of the significance test is given, meaning that if its value is smaller than a significance threshold (0.10 was used) the contribution of the predictor to the adjustment of the model is considered significant. Model quality is presented by means of the confusion matrix.

The comparison between the binary logistic model and the three alternative classification methods is also presented by means of three summary tables for the 5-fold crossvalidation method:

i) the overall correct classification rates for all the classification methods, for each crossvalidation fold, and the correspondent average values;
ii) the specificity classification rates for all the classification methods, for each crossvalidation fold, and the correspondent average values, and
iii) the sensitivity classification rates for all the classification methods, for each crossvalidation fold, and the correspondent average values.
The statistical analysis was performed using XLSTAT-Pro, Version 05.33993, 2016, Addinsoft, Inc., USA.

The binary logistic model for student dropout

The goodness-of-fit statistics presented in Table 3 show that the model is an overall good representation of the relationship between student dropout and students' success/failure in the 12 courses of the first year. In fact, $R^2_{McF} = 0.502$ reveals that the model is well adjusted, and the likelihood ratio (χ^2) of 134.57[5] means that the adjusted model is significantly more powerful than the independent model (which gives probability p_0 whatever the values of the predictors are).

5 - $P(\chi^2_{GL=4})>134.57)<0.0001$

Table 3- Goodness-of-fit statistics for the Logistic regression model

Statistics	Intercept-only model	Full model	Likelihood ratio (χ^2)	DF	Pr > χ^2
-2 Log Likelihood	267.860	133.290	134.57	4	<0.0001
R^2(McFadden)	0.000	0.502	-	-	-

Source: Authors.

Only four of the 17 potential predictors were included in the model. The significance of the four included predictors is presented in Table 4.

Table 4- Significance analysis for the predictors

Predictor	Predictor/Acronyms	Type	Pr > LR
X'_7	Materials Science (Mat_Sc)	Categorical (0-pass / 1-fail)	<0.001
X'_{14}	Electricity (Elect)	Categorical (0-pass / 1-fail)	<0.003
X'_{12}	Calculus 1 (Cal_1)	Categorical (0-pass / 1-fail)	<0.059
X'_{13}	Chemistry (Che)	Categorical (0-pass / 1-fail)	<0.078

Source: Authors.

Equation 2 shows the expression for the calculation of the probability of a student being in class 1 (a student who dropped out), conditional to the predictors *X'*.

Equation (2)

$$P(Class=1 \mid X') = \frac{1}{1+e^{-(-2.964+2.058*Mat\_Sc+1.430*Elect+0.927*Cal\_1+1.116*Che)}}$$

In equation 2, when all the four predictors (Mat_Sc, Elect, Cal_1, and Che marks) equal zero (pass the course) the probability of dropout is

Equation (3) $P(Class=1 \mid X'=0)=0.049$

When changing each of the predictors from 0 to 1 ("pass" the course to "fail" in the course), maintaining the remaining three predictors' values as zero, the probabilities of dropout are as follows:

Equation (4) $P(Class=1 \mid \text{Mat\_Sc}=1,\text{Elect}=0,\text{Cal\_1}=0,\text{Che}=0)=0.288$

Equation (5) $P(Class=1 \mid \text{Elect}=1,\text{Mat\_Sc}=0,\text{Cal\_1}=0,\text{Che}=0)=0.177$

Equation (6) $P(Class=1 \mid \text{Cal\_1}=1,\text{Mat\_Sc}=0,\text{Elect}=0,\text{Che}=0)=0.115$

Equation (7) $P(Class=1 \mid \text{Che}=1,\text{Mat\_Sc}=0,\text{Elect}=0,\text{Cal\_1}=0)=0.136$

In equation 2, when all the four predictors (Mat_Sc, Elect, Cal_1, and Che marks) equal one (fail in the course) the probability of dropout is

Equation (8) $P(Class=1 \mid \text{X'}=1)=0.929$

The predictive model consistency and accuracy results were assessed using cross-validation – 5 folds. Table 5 presents the summaries for accuracy (the proportion of the total number of predictions that were correct), for sensitivity (the proportion of actual positive cases that were classified correctly) and for specificity (the proportion of actual negatives cases that were classified correctly) obtained by running the logistic model using the validation sample as input data. These summary results show that the use of the binary logistic regression model produced very good overall accuracy in classifying the students who dropped out.

Table 5- Predictive accuracy results for the binary Logistic model, using as input the validation sample: Cross-validation summary results

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVERAGE
Accuracy	82%	85%	84%	90%	84%	85%
Specificity	90%	86%	86%	90%	76%	86%
Sensivity	71%	83%	82%	89%	94%	84%

Source: Authors.

Results for other classification methods

The results for the following classification models were obtained by running each model using the validation sample as input data.

a) One R method

This classification rule is based on the value of a single predictor. On the 5-fold cross-validation, two of the four predictors used in the binary logistic model (see Table4) have been selected in different folds as the single classifier for student dropout: Electricity results (0 - pass the course / 1 - fail in the course) were selected in three of the five folds and Materials Science results (0 - pass the course / 1 – fail in the course) were selected in two of the five folds. Table 6 presents the summaries for accuracy, for sensitivity and for specificity obtained by running the One R methodology using the validation sample as input data.

Table 6- Predictive model accuracy results for the One R classification method: Cross-validation summary results

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVERAGE
Accuracy	68%	63%	58%	61%	61%	62%
Specificity	90%	95%	86%	90%	86%	89%
Sensivity	41%	28%	24%	28%	29%	30%

Source: Authors.

b) K Nearest Neighbors, using K=7

This methodology classifies the input data based on a distance computed using the four predictors used in the binary logistic model (see Table4). Table 7 presents the summaries for accuracy, for sensitivity and for specificity obtained by running the K Nearest Neighbors methodology using the validation sample as input data. We tried several values for K in the range from 3 to 10 and select K=7 since it was the value that produced better results.

Table 7- Predictive model accuracy results for the K Nearest Neighbors classification method: Cross-validation summary results

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVERAGE
Accuracy	82%	84%	84%	87%	82%	84%
Specificity	90%	85%	86%	85%	76%	84%
Sensivity	71%	83%	83%	89%	88%	83%

Source: Authors.

c) Naive Bayes

This methodology classifies the input data based on Bayes' Theorem with strong independence assumptions. The methodology uses the four predictors that have been used in the binary logistic model (see Table4). Table 8 presents the summaries for accuracy, for sensitivity and for specificity obtained by running the Naive Bayes model using the validation sample as input data.

Table 8- Predictive model accuracy results for the Naive Bayes classification method: Cross-validation summary results

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVERAGE
Accuracy	82%	89%	84%	92%	84%	86%
Specificity	90%	95%	90%	91%	81%	90%
Sensivity	71%	83%	76%	94%	88%	82%

Source: Authors.

Discussion and conclusions

As a case study, the analysis carried out in this research was intended to be transposable to practical actions that may reduce the likelihood of voluntary withdrawal from a specific program of the ISVOUGA Institute. The results obtained are not easily comparable to results obtained by other researchers in the same area since most of their studies (at least the ones that we are aware of) basically try to explain dropout using variables such as the teaching staff, the student body, the institutional, social and family contexts (ARAQUEA; ROLDÁN; SALGUERO, 2009).

In this research we examined the effect of several intrinsic factors (student-related) and several extrinsic factors (institutional-related) on the dropout of students enrolled in the Industrial Engineering program of the ISVOUGA Institute. A dataset containing data from 192 students consisting of students' general information (selected course, name, gender, marital status, professional status, full-time student or part-time student, age at the beginning of the program) and students' marks was used to produce a set of predictors for students' dropout. Seventeen variables have been considered as potential predictors (see Table 2) representing student-related and institutional-related dropout influent factors.

We performed a binary logistic regression that classified students as 0 (student who did not drop out of the program) or 1 (student who dropped out of the program), using the probability that the student belongs to each class and a probability threshold of 0.5 to separate both classes. Moreover, three classification methods commonly used in Data Mining have been employed to compare their overall level of accuracy with the one obtained by using binary logistic regression.

Binary logistic regression is a classical statistics method used to classify/predict the likelihood of an item belonging to one of two classes (it is a method still used in Data Mining). Several other Data Mining classification methods could be used instead of a binary logistic regression (we explored three of these methods: One-R, K Nearest Neighbors, and Naive Bayes). In this research, the binary logistic regression model has produced a very good overall accuracy level (see Table 5), comparable to the accuracy level of the Naive Bayes method but superior to the accuracy level of One-R or K Nearest Neighbors methods. However, some authors criticize Data Mining techniques for their long training process, their inability to identify the relative importance of the potential input predictors, and their interpretative difficulties (SABZEVARI; SOLEYMANI; NOORBAKHSH, 2007). The binary logistic model has produced high quality predictions, using a reduced number of predictors whose selection was based on their significance. This fact facilitates the understanding the contribution of each predictor to the adjustment of the model. Additionally, models' equation 2 allows the assessment of the level of influence that each predictor has on models' output.

Four of the 17 available variables were considered significant to the model and included as predictors. These predictors are the categorical results (0 - pass the course / 1 - fail in the course) obtained by students in four courses taught during the first two semesters of the Industrial Engineering program: Materials Science, Electricity, Calculus 1, and Chemistry. The first two predictors included in the logistic regression model,

Materials Science and Electricity, are substantially more influent in model output than Calculus 1 and Chemistry. It is interesting to note that the two most influential predictors for student dropout are failing in the less challenging courses of Materials Science (level of difficulty 2) and Electricity (level of difficulty 2) – see the level of difficulty of all courses in Table 1 (column 5). Failing in the very challenging courses of Calculus 1 (level of difficulty 4) and Chemistry (level of difficulty 4) is less influential in student dropout than failing in the Materials Science or Electricity courses. Moreover, neither of the variables X_1 – Gender, X_2 – Marital status, X_3 – Professional status, X_4 – Part/full time student nor X_5 – Age at the beginning of the program, have been included in the model as they were not statistically significant. When, in equation 2, all the four predictors equal one (fail in the course) the probability of dropout is 0.929 being 0.049 when all the four predictors equal zero (pass the course) (see equations (3) and (8)). The logistic model showed an overall accuracy (the proportion of the total number of predictions that were correct) of 85% when using the validation samples as input (see Table 5).

Contrary to what we would think to be expectable prior to this research, we found that to be retained (failing) in more challenging courses such as Physics (difficulty 4 in 1-4 scale, see Table 1), Spreadsheets (MS Excel), VB Programing, and Statistics (difficulty 3 in 1-4 scale, see Table 1) does not have significant influence on student dropout. Surprisingly, being retained on less challenging courses such as Materials Science (difficulty 2 in a 1-4 scale – see Table 1) and Electricity (difficulty 2 in 1-4 scale, see Table 1) – has significant influence on student dropout. The reason for this ambiguity is not clearly determinable by the analysis of the results of this research although it seems to us reasonable to hypothesize that when freshmen seek information about the Industrial Engineering program, prior to their registration, they interiorize that several courses within the program do have a high level of difficulty. This prior knowledge makes freshmen intrinsically prepared to be retained on some of these courses. Conversely, failing in courses that freshmen would consider to be of low level of difficulty will demotivate them and in some cases will drive them to dropout.

The findings of this research will be used at ISVOUGA for the adjustment of the syllabus of the Materials Science and Electricity courses. Moreover, these courses will be repositioned across the 6 semesters of the Industrial Engineering program in such a way that may improve retention and success rates, keeping students motivated to complete their program of studies. Additionally, peer mentoring, using successful students to model studying habits for the freshmen, should be tried in order to increase the pass rates and reduce the overall dropout rates.

The same type of analysis used in this research regarding the Industrial Engineering program of ISVOUGA may be of use for any other institution in order to identify the courses that do not comply with the expectations of students in terms of level of effort to pass. In this research results, it appears that failing in less challenging courses of a program is more demotivating to students than failing in very challenging ones. In some situations, it will be possible to adjust the syllabus of these courses so that, without compromising their overall purpose, they may become more tied to the students' expectations. When these

syllabus adjustments are not an effective possibility then, if possible, these courses should be repositioned over the program in a such a way that by the time students attend them they will be better prepared to their intrinsic level of difficulty, putting in the necessary effort to pass them.

References

ARAQUE, Francisco; ROLDÁN, Concepcion; SALGUERO, Alberto. Factors influencing university drop out rates. **Computers and Education**, Amsterdã, v. 53, n. 3, p. 563-574, 2009.

ARWU. **Academic Ranking of World Universities 2016.** Disponível em: <http://www.shanghairanking.com/ARWU2016.html>. Acesso em: 13 out. 2016.

ASTIN, Alexander. Student involvement: a developmental theory for higher education. **Journal of College Student Development**, Baltimore, v. 40, n. 5, p. 518-529, 1984.

ASTIN, Alexander. **What matters in college:** four critical years revisited. San Francisco: Jossey-Bass, 1993.

BOLOGNA WORKING GROUP. **A framework for qualifications of the European higher education area.** Copenhagen: Ministry of Science, Technology and Innovation, 2005.

COVER, Thomas; HEART, P. Nearest neighbor pattern classification. **Information Theory**, Piscataway, v. 13, n. 1, p. 21-27, 1967. DOI: 10.1109/TIT.1967.1053964.

HAND, David J. Principles of data mining. **Drug Safety**, Berlim, v. 30, n. 7, p. 621-622, 2007.

HOVDHAUGEN, Elisabeth. Do structured study programmes lead to lower rates of dropout and student transfer from university? **Irish Educational Studies**, London, v. 30, n. 2, p. 237-251, 2011. Disponível em: <https://doi.org/10.1080/03323315.2011.569143>. Acesso em: 13 out. 2106.

INE. Instituto Nacional de Estatística. **Censos 2011 norte**. v. 60. [S.l.] INE, 2011.

INE. Instituto Nacional de Estatística. **Portal do Instituto Nacional de Estatística.** [S.l.] INE, 2016. Disponível em: <https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0003920&contexto=bd&selTab=tab2>. Acesso em: 30 out. 2016.

JORDAN, Will J.; LARA, Julia; McPARTLAND, James M. **Exploring the complexity of early dropout causal structures.** Baltimore: Center for Research on Effective Schooling for Disadvantaged Students, 1994.

KUH, George D. What we're learning about student engagement from NSSE: benchmarks for effective educational practices. Change. **The Magazine of Higher Learning**, London, v. 35, n. 2, p. 24-32, 2003. DOI: https://doi.org/10.1080/00091380309604090.

LITALIEN, David; GUAY, Frédéric. Dropout intentions in PhD studies: a comprehensive model based on interpersonal relationships and motivational resources. **Contemporary Educational Psychology**, Amsterdã, v. 41, p. 218-231, Apr. 2015. DOI: https://doi.org/10.1016/j.cedpsych.2015.03.004.

LONG, J. Scott. Regression models for categorical and limited dependent variables. In: PAPER SERIES on quantitative applications in the social sciences. Thousand Oaks: Sage, 1997. p. 52-54. (Series n. 07, 7).

MALM, Joakim; BRYNGFORS, Leif; MÖRNER, Lisse-Lotte. Supplemental instruction for improving first year results in engineering studies. **Studies in Higher Education**, London, v. 37, n. 6, p. 655-666, 2012. DOI: https://doi.org/10.1080/03075079.2010.535610.

MALM, Joakim; BRYNGFORS, Leif; MÖRNER, Lise-Lotte. The potential of supplemental instruction in engineering education - helping new students to adjust to and succeed in university studies. **European Journal of Engineering Education**, London, v. 40, n. 4, p. 347-365, 2015. DOI: https://doi.org/10.1080/03043797.2014.967179.

MARTÍNEZ-LÓPEZ, Zelita et al. Apoyo social en universitarios españoles de primer año: propiedades psicométricas del social support questionnaire-short form y el social provisions scale. **Revista Latinoamericana de Psicologia**, Barcelona, v. 46, n. 2, p. 102-110, 2014. DOI: https://doi.org/10.1016/S0120-0534(14)70013-5.

MENARD, Scott. Applied logistic regression analysis. In: PAPER SERIES on quantitative applications in the social sciences. 2. ed. Thousand Oaks: Sage, 2001. p. 75-79. (Series n. 07-106).

MÍNGUEZ, Almudena Moreno; SAN JULIÁN, Elena Rordríguez. **Informe juventud en España 2012.** Madrid: [s.n., 2012].

MORALES, Erik E.; AMBROSE-ROMAN, Sarah; PEREZ-MALDONADO, Rosa. Transmitting success: comprehensive peer mentoring for At-Risk students in developmental math. **Innovative Higher Education**, Madrid, v. 41, n. 2, p. 121-135, 2016.

NEILD, Ruth Curran; BALFANZ, Robert; HERZOG, Liza. An early warning system. **Educational Leadership**, Alexandria, v. 65, n. 2, p. 28-33, 2007.

OECD. Portugal. In: EDUCATION at a glance 2017: OECD indicators. Paris: OECD, 2017. p. 42-50.

PÁRAMO FERNÁNDEZ, Maria Fernanda et al. Predictors of students' adjustment during transition to university in Spain. **Psicothema**, Bethesda, v. 29, n. 1, p. 67-72, fev. 2017. DOI: 10.7334/psicothema2016.40.

PASCARELLA, Ernest T.; TERENZINI, Patrick T. Predicting freshman persistence voluntary dropout decisions from a theoretical model. **The Journal of Higher Education**, New York, v. 51, n. 1, p. 60-75, 1980. DOI: 10.2307/1981125.

PAURA, Liga; ARHIPOVA, Irina. Cause analysis of students' dropout rate in higher education study program. In: WORLD CONFERENCE ON BUSINESS, ECONOMICS AND MANAGEMENT, 2., 2014, Amsterdã. **Anais…** v. 109. Amsterdã: [s. n.], 2014. p. 1282-1286. DOI: https://doi.org/10.1016/j.sbspro.2013.12.625.

PIERRAKEAS, Christos et al. A comparative study of dropout rates and causes for two different distance education courses. **International Review of Research in Open and Distance Learning**, Athabasca, v. 5, n. 2, p. 1-14, 2004.

REFAEILZADEH, Payam; TANG, Lei; LIU, Huan. Cross-validation. In: ENCYCLOPEDIA of database systems. [S. l.]: Springer, 2009. p. 532-538.

RUSSEL, Stuart; NORVIG, Peter. Artificial intelligence: a modern approach. New Jersey: Prentice Hall, 1995.
SABZEVARI, Hassan; SOLEYMANI, Mehdi; NOORBAKHSH, Eman. A comparison between statistical and data mining methods for credit scoring in case of limited available data. In: CRC CREDIT SCORING CONFERENCE, 3., 2007, Edinburgh. **Proceedings of the…** Edinburgh: [s. n.], 2007. p. 1-8.

STRATTON, Leslie Stundt; O'TOOLE, Dennis M.; WETZEL, James N. Are the factors affecting dropout behavior related to initial enrollment intensity for college undergraduates? **Research in Higher Education**, Berlim, v. 48, n. 4, p. 453-485, Feb. 2007.

TINTO, Vincent. Limits of theory and practice in student attrition. **The Journal of Higher Education**, New York, v. 53, n. 6, p. 687-700, 1982. DOI: 10.2307/1981525.

WILSON, D. Randall; MARTINEZ, Tony R. Improved heterogeneous distance functions. **Journal of Artificial Intelligence Research**, v. 6, p. 1-34, 1997.

Received on June 2nd, 2017
Revisions on October 25th, 2017
Approved on January 21st, 2018

António Carlos Corte-Real de Sousa is an assistant professor at Instituto Superior de entre Douro e Vouga (ISVOUGA), visiting assistant professor at the Departamento de Engenharia e Gestão Industrial da Faculdade de Engenharia da Universidade do Porto (FEUP), integrated member of the Centro de Engenharia e Gestão Industrial (CEGI) of INESC TEC.

Carlos Alberto Bragança de Oliveira is an assistant professor at Departamento de Engenharia e Gestão Industrial da Faculdade de Engenharia da Universidade do Porto (FEUP), Member of Conselho de Departamento de Engenharia e Gestão Industrial, integrated member of the Centro de Engenharia e Gestão Industrial (CEGI) do INESC TEC.

José Luís Cabral Moura Borges is an associate professor at Departamento de Engenharia e Gestão Industrial da Faculdade de Engenharia da Universidade do Porto (FEUP), member of Conselho de Departamento de Engenharia e Gestão Industrial, integrated member of the Centro de Engenharia e Gestão Industrial (CEGI) do INESC TEC, collaborator of Interface Institute: Instituto de Ciência e Inovação em Engenharia Mecânica e Engenharia Industrial.