A Data Mining Approach for Predicting Academic Success – A Case Study: Helping Teachers Develop Research Informed Practice

Chapter · February 2019

DOI: 10.1007/978-3-030-11890-7_5

CITATIONS
4

READS
790

4 authors, including:

Maria Martins
Instituto Politécnico de Bragança
10 PUBLICATIONS **20** CITATIONS
SEE PROFILE

V.L. Miguéis
University of Porto
43 PUBLICATIONS **1,065** CITATIONS
SEE PROFILE

Davide S. B. Fonseca
Universidade da Beira Interior
65 PUBLICATIONS **559** CITATIONS
SEE PROFILE

A Data Mining Approach for Predicting Academic Success – A Case Study

Maria P. G. Martins[1,3(✉)], Vera L. Miguéis[2], D. S. B. Fonseca[3], and Albano Alves[1]

[1] School of Technology and Management, Polytechnic Institute of Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal
prud@ipb.pt
[2] Faculty of Engineering, University of Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[3] CISE - Electromechatronic Systems Research Centre, University of Beira Interior, Calçada Fonte do Lameiro, P, 6201-001 Covilhã, Portugal

Abstract. The present study puts forward a regression analytic model based on the random forest algorithm, developed to predict, at an early stage, the global academic performance of the undergraduates of a polytechnic higher education institution. The study targets the universe of an institution composed of 5 schools rather than following the usual procedure of delimiting the prediction to one single specific degree course. Hence, we intend to provide the institution with one single tool capable of including the heterogeneity of the universe of students as well as educational dynamics. A different approach to feature selection is proposed, which enables to completely exclude categories of predictive variables, making the model useful for scenarios in which not all categories of data considered are collected. The introduced model can be used at a central level by the decision-makers who are entitled to design actions to mitigate academic failure.

Keywords: Data mining · Educational data mining · Prediction · Academic success · Random forest · Regression

1 Introduction

The quality of academic training has a paramount role in the growth and development of any country or society. In turn, educational success is closely linked to the efficacy and efficiency of educational institutions. For this reason, the delivery of high-quality training and the definition of strategies which may promote academic success as well as retention recovery have been the subject of deep reflection by the administration board of the Polytechnic Institute of Bragança (IPB – *Instituto Politécnico de Bragança*), a polytechnic higher education institution in inland Portugal. Therefore, the aim to provide a methodology which enables the obtainment of useful knowledge to help and ground decision-making

Á. Rocha et al. (Eds.): ICITS 2019, AISC 918, pp. 45–56, 2019.
https://doi.org/10.1007/978-3-030-11890-7_5

by the IPB management bodies created the need to develop an academic success predictive model, which fits educational data mining field.

This work presents a regression model based on the random forest algorithm [1], developed with the aim to predict at an early stage the global academic performance of IPB undergraduates at the terminus of their academic path.

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 of results and of the prediction model of performance proposed; Sect. 5 – final discussion of results and respective conclusions.

2 Related Studies

The main goal of educational data mining (EDM) is to generate useful knowledge which may ground and sustain decision-making targeted at improving student communities' learning as well as educational institutions' efficiency. Several systematic literature reviews [2–9] give evidence of the growing importance of EDM throughout time and refer and analyse the main research topics in which EDM has shown a remarkable contribution as a management analysis and support tool. Also, such studies provide evidence of the usefulness, potential and efficacy of the most used data mining methods and algorithms.

Among the typology of tasks where EDM has shown a remarkable contribution, we can find the prediction of academic performance. This research topic is normally approached from three different perspectives: predicting school dropout, predicting retention, or predicting academic success at the end of the degree course. Literature on EDM has shown that within those studies, several authors over time have studied a number of factors which promote academic success or failure. The main goal of the studies [10,11] was to conduct a review on the most used and relevant factors for this kind of predictions. Simultaneously, the authors also intended to determine the main data mining methods and algorithms used in such studies. After analysing a set of 30 studies focusing on the topic, Shahiri et al. [10] conclude that 6 attributes are used most frequently. At the top of the list is the cumulative grade points average (CGPA), almost at the same level as the attributes of internal assessment (marks after entering higher education such as assignments, exams, attendance, etc.). Demographic features and external assessment (pre-university achievement classifications) were the second group of most used attributes. Finally, the third group of attributes that the authors considered to be most used among the set of 6 are the ones related to students' extra-curricular activities and social interaction.

In a similar study to that by Shahiri et al. [10], but with the particularity of focusing only on research related to students attending institutions of the traditional on-site system, Del Río and Insuasti [11] conclude that in order to infer the final mean of the degree course, the authors of a set of 51 studies released between 2011 and August 2016 used, as predictive variables, indicators of academic performance obtained after entering higher education in combination with another type of attribute in 51.8% of the studies. In 37.5% of the

papers, they only used information on academic performance within higher education. Among the data mining methods most used in the EDM task, the same authors [11] highlight those regarding classification, reported in 71.4% of the research works mentioned. The methods which followed were those of clustering and association rules, present in 8.9% and 7.1% of the studies.

Regarding the predicting academic success at the end of the degree course, the majority of the works related analyse academic performance restricted to one degree course only, and use relatively small datasets. For example, Natek and Zwilling [12] concluded that the factors most visibly influencing the final mean of the undergraduates of the bachelor degree in Computer Science were connected to information regarding access, demography and extra-curricular activities. The data were processed by the classification algorithms RepTree Model, J48 Model and M5P Model, and concerned 42 students attending the 1st year of the degree course, 32 attending the 2nd year, and 32 attending the 3rd year.

The results of the study conducted by Asif, Merceron, Ali and Haider [13] show that through algorithms Naïve Bayes and random forest Trees, it is possible to predict, with a high level of precision, the global graduation performance of a four-year degree course, by using only pre-university marks and the marks obtained in the course units of the 1st and 2nd years of university.

In the study by Migueis et al. [14], random forest, decision trees, support vector machines, Naïve Bayes, bagged trees and boosted trees were used to predict overall students academic performance based on the information available at the end of the first academic year. Among the algorithms used, random forest was the one that showed the best predictive results, also providing evidence that the most important factors to predict and explain the level of academic success in five-year degree courses in Engineering are the means regarding university access and university access examinations, as well as the mean obtained in the course units of the first academic year. This study also proposes a multiclass segmentation structure, aiming an early classification of students, based on their performance observed at the end of the first academic year and on their propensity for academic success revealed by the predictive model.

3 Methodology and Data Model

For the creation of a predictive model which can predict students' academic success at the terminus of their academic path, we chose to explore data from a universe of students from different educational fields, attending about half a hundred degree courses in an institution made up of 5 schools instead of following the most common procedure of delimiting the prediction to one single specific course. Hence, we intend to provide the institution with one single tool capable of including the heterogeneity of the universe of students as well as educational dynamics. The aim is for this tool to be used at a central level by the decision-makers who are entitled to design actions to mitigate academic failure, thus promoting a better educational experience for their students.

In this study, we chose to base our predictive model on the random forest algorithm proposed by Breiman [1]. It has shown to have surpassed other techniques in similar studies due to its predictive capacity and, more importantly, it allows a good interpretation of its results, in contrast to other techniques, such as Neural Networks and Support Vector Machines, which are considered to be black boxes. In fact, random forest presents the interesting functionality of allowing ordering the importance of the predictors which sustain the model.

Following the procedure commonly adopted, the predictive performance of the models considered was evaluated using a *k-fold* (with $k = 10$) cross-validation. As for the model evaluation metrics, the determination coefficient (R^2) and the root mean squared error (RMSE) were used.

In order to determine students' academic performance, the dependent variable introduced in expression (1) was used as a success indicator,

$$dv = average \times \frac{ects\_aprov}{ects\_aprov + ects\_disaprov}, \tag{1}$$

where `average` is the weighted average of the marks obtained in the completed course units (CUs), `ects_aprov` is the number of ECTS completed successfully and `ects_disaprov` is the number of ECTS in which students enrolled but did not pass. Thus, the metric takes into account not only the students' classification average but also the fraction of matriculations in course units they passed (ratio of successful 'attempts').

It was possible to consider the time period comprised between 2007/2008 and 2015/2016, totalising 9 consecutive academic years. A choice was made to limit the study to the bachelor degree courses since they are the core of the institution's training offer and they encompass a more complete set of data. After a cleaning of data and other pre-processing tasks, the data set that this study focuses on comprised 4530 matriculations in bachelor degree courses concluded in the period between 2007/2008 and 2015/2016 and started in the period between 2007/2008 and 2013/2014.

Regarding the predictive variables of academic success, we considered essentially the same typology of variables used in the related works of reference, namely academic data of sociodemographic nature and of access to higher education. The variables under study can be classified in two important subgroups: variables with cumulative semestral curricular results and 'timeless' variables – variables whose values are unaltered throughout students' academic path. Table 1 presents all the 44 predictive variables considered in this study as well as the dependent variable `dv` which will be used as a success indicator according to Eq. (1).

A vast real data set was used in this study, so care was taken to classify (3rd column of the table) each one of the potential predictive variables of academic success according to their nature into five different categories: *curricular* (C), *matriculation* (M), *demographic* (D), *socioeconomic* (S) and *access* (A). The attributes regarding semestral data (all in category C) are also easily distinguished from the others (timeless data) through the suffix '_s'. Note that the

Table 1. List of variables sustaining the model.

Id	Attribute	Cat	Type	Min..max	Meaning
1	curricular_year_s	C	Discrete	1..4	Student's course year in the a.s. considered
2	academic_year_s	C	Discrete	07..15	Academic year of the a.s. considered
3	scholarship_s	C	Continuous	0..1	Was the student a scholarship holder in the a.s.?
4	union_member_s	C	Continuous	0..1	Was the student a union leader in the a.s.?
5	ects_aprov_s	C	Discrete	0..60	N. of ECTS passed in the a.s.
6	ects_disaprov_s	C	Discrete	0..60	N. of ECTS failed in the a.s. (academic semester)
7	max_s	C	Discrete	0..20	Maximum mark of the CUs passed in the a.s.
8	average_s	C	Continuous	0..20	Average mark of the CUs passed in the a.s.
9	min_s	C	Discrete	0..20	Minimum mark of the CUs passed in the a.s.
10	n_assess_disap_s	C	Discrete	0..18	N. of assessments failed in the a.s.
11	n_courses_aprov_s	C	Discrete	0..10	N. de CUs passed in the academic semester
12	n_courses_disap_s	C	Discrete	0..10	N. of CUs failed in the academic semester
13	dv12_s[(a)]	C	Continuous	−20..20	Diff. in performance from 1st to 2nd semester
14	dv23_s[(a)]	C	Continuous	−20..20	Diff. in performance from 2nd to 3rd semester
15	dv34_s[(a)]	C	Continuous	−20..20	Diff. in performance from 3rd to 4th semester
16	dv45_s[(a)]	C	Continuous	−20..20	Diff. in performance from 4th to 5th semester
17	dv56_s[(a)]	C	Continuous	−20..20	Diff. in performance from 5th to 6th semester
18	enrol_year	M	Discrete	07..13	Year of enrolment
19	cod_degree	M	Nominal	1..51	Code of the degree course
20	cod_school	M	Nominal	1..5	Code of the school
21	cred_ects_tx	M	Discrete	0..100	Fraction of ECTS credited to the student
22	ects_degree	M	Discrete	180..240	Number of ECTS of the degree course
23	enrol_type	M	Nominal	1..9	Type of enrolment in the degree course
24	displaced	D	Binary	0..1	Is the student displaced from usual residence?
25	district	D	Nominal	1..28	Student's district of origin
26	district_n	D	Nominal	1..27	District of birth
27	age	D	Discrete	17..61	Student's age at the time of enrolment
28	nationality	D	Nominal	1..15	Student's nationality
29	gender	D	Nominal	1..2	Gender
30	cod_job_student	S	Nominal	1..12	Student's job
31	cod_job_mother	S	Nominal	1..12	Mother's job
32	cod_job_father	S	Nominal	1..12	Father's job
33	educ_level_mother	S	Ordinal	1..13	Mother's level of education
34	educ_level_father	S	Ordinal	1..13	Father's level of education
35	prof_sit_student	S	Nominal	1..10	Student's employment status
36	prof_sit_mother	S	Nominal	1..10	Mother's employment status
37	prof_sit_father	S	Nominal	1..9	Father's employment status
38	phase	A	Ordinal	1..3	Phase of enrolment
39	access_grade	A	Continuous	0..200	Student's entrance qualification
40	a10_11_grade	A	Continuous	0..200	Mean obtained in the 10th and 11th grades
41	a12_grade	A	Continuous	0..200	Mean obtained in the 12th grade
42	access_option	A	Ordinal	1..6	Order of the option when applying for university
43	access_order	A	Discrete	1..322	Order of entrance among stud. admit. in course
44	access_exams	A	Continuous	0..200	Mean obtained in the entrance examinations
45	dv[(b)]		Continuous	0..20	Dependent var. with student's final performance

[(a)] $dv_{ij}\_s = dv_j - dv_i$, with $i = j - 1$ and dv_n the student's performance in their umpteenth semester, calculated using a metrics similar to that of dv (c.f. Eq. (1)). Each variable dv_{ij}_s is only present in the models with cumulative data of j or more semesters.
[(b)] Dependent variable used as indicator of success.

curricular data (C) refers to semestral curricular results of academic performance accumulated at the end of each of the student's 6 first semesters.

The selection of the dimensions regarding the student, which explain their academic success, was carried out in two different stages. First, as explained in Sect. 4.1, we selected the student's dimensions which best account for their success, which allowed a first adjustment enabling the exclusion of complete groups of variables. Subsequently, as explained in Sect. 4.2, we fine-tuned the selection of attributes which were not excluded in the first stage. This approach allowed the reduction of the data dimensionality without losing the model's predictive capacity.

Figure 1 depicts a scheme intended to characterise the predictive model designed for this study. It shows the different categories of predictive variables used as input of the random forest algorithm. As we can see, for the group of curricular variables (C), the attributes used are the results accumulated at the end of each one of the student's 6 first semesters. Note that only one of the 6 entries of curricular data (C) is considered in each execution of the algorithm (mutually exclusive entries).

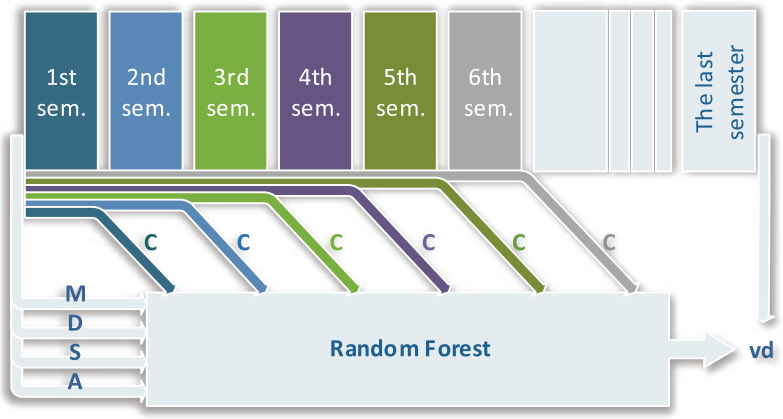

Fig. 1. Scheme depicting the comparative study conducted.

4 Results

In the exploratory analysis of data which follows, we chose to keep the configuration of the random forest algorithm fixed for the study to focus on the set of predictive variables which sustain it.

4.1 Selection of Predictors Categories

As it is known, the assertiveness of a predictive model depends greatly on the set of predictive variables being considered in the analysis. Also, the best model is not always the one which includes all the variables available. Although it

is commonly acknowledged that random forests make an internal selection of variables, we still decided to run a test of inclusion, or not, of the categories of variables so as to understand the impact of the several dimensions on the capacity of the model.

Table 2 shows the different categories of attributes chosen in each of those studies. Although the aim is to develop a comprehensive study, it does not seem necessary to include all the possible combinations between the 5 groups of variables, which makes a total of $2^5 - 1 = 31$ possibilities. In fact, since the group of curricular data (C) is clearly the most determinant group of predictors, great difficulties are foreseen in the predictive precision of any model which does not include it. Therefore, only one particular case is considered in which this group of variables is not used: case MDSA (Study 6). With this simplification, the number of studies was reduced to almost a half, or more precisely to $2^4 + 1 = 17$.

Table 2. Categories of predictive variables used in the studies conducted.

Study	Label	**C**urricular	**M**atriculation	**D**emographic	**S**ocioeconomic	**A**ccess
1	CMDSA	✓	✓	✓	✓	✓
2	CMDS	✓	✓	✓	✓	
3	CMDA	✓	✓	✓		✓
4	CMSA	✓	✓		✓	✓
5	CDSA	✓		✓	✓	✓
6	MDSA		✓	✓	✓	✓
7	CMD	✓	✓	✓		
8	CMS	✓	✓		✓	
9	CMA	✓	✓			✓
10	CDS	✓		✓	✓	
11	CDA	✓		✓		✓
12	CSA	✓			✓	✓
13	CM	✓	✓			
14	CD	✓		✓		
15	CS	✓			✓	
16	CA	✓				✓
17	C	✓				

Table 3 contains for each of the studies the determination coefficients (R^2) obtained by applying the random forest predictive algorithm to the selected data at the end of the student's 6 first academic semesters. For a better understanding of the results presented, take Study 8 as an example: the input of the random forest algorithm (see Fig. 1) was reduced to the groups of curricular (C), matriculation (M) and socioeconomic (S) variables. In this study, as in all the others (except for Study 6, which does not include the curricular data), the random forest algorithm was run 6 times in order to use the results accumulated at the end of each of the 6 semesters concerning the curricular data. The last column of the table shows the weighted average of the determination coefficients for those

Table 3. Determination coefficient R^2 of the predictive model, for different groups of predictive variables according to students' academic semester.

	Label	1st sem	2nd sem	3rd sem	4th sem	5th sem	6th sem	Average[(a)]
Study 13	CM	80.4	86.5	92.0	94.3	96.6	97.9	88.4
Study 8	CMS	80.7	86.8	91.7	93.9	96.4	97.7	88.4
Study 4	CMSA	80.7	86.5	91.6	93.9	96.3	97.7	88.3
Study 1	CMDSA	80.3	86.3	91.3	93.7	96.2	97.6	88.1
Study 2	CMDS	80.3	86.4	91.4	93.7	96.3	97.7	88.1
Study 7	CMD	79.8	86.3	91.6	94.0	96.5	97.8	88.1
Study 9	CMA	79.9	86.2	91.6	94.0	96.4	97.7	88.1
Study 3	CMDA	79.7	86.1	91.4	93.8	96.3	97.7	87.9
Study 12	CSA	70.5	78.8	86.6	89.9	94.2	96.6	81.8
Study 15	CS	70.6	78.7	86.5	89.7	94.1	96.5	81.8
Study 5	CDSA	70.5	78.5	86.4	89.8	94.2	96.5	81.7
Study 10	CDS	70.7	78.3	86.2	89.6	94.1	96.4	81.6
Study 17	C	70.3	78.0	86.6	90.0	94.4	96.7	81.6
Study 16	CA	69.6	78.2	86.6	90.1	94.4	96.7	81.5
Study 11	CDA	69.8	78.2	86.3	89.9	94.3	96.6	81.4
Study 14	CD	70.1	77.8	86.3	89.6	94.2	96.6	81.4
Study 6	MDSA	64.4	64.4	64.4	64.4	64.4	64.4	64.4
Average[(b)]		75.2	82.4	89.0	91.9	95.3	97.2	84.9

[(a)] Weighted average of semestral values, with weights 6, 5, ..., 2, 1, for the 1st, 2nd, ..., 5th, 6th semesters, respectively.
[(b)] Average value without considering Study 6.

6 semesters. The choice fell on a weighted average of semestral R^2 in order to value the results of the first semesters at the expense of those obtained in more advanced stages of the student's academic path – note, for example, that for students who complete their training in 3 years, the predictive capacity of the model after the 6th semester is totally irrelevant.

The analysis of the values presented in the table, listed in a descending order of the average value of R^2 (last column), allows the following considerations:

- It was not the model 'feeding' on all the variables (Study 1 – CMDSA) which presented the best predictive capacities. Actually, 6 other models achieved similar or better performances with a lower number of predictive variables.
- There is a visibly big difference in the performance of the 8 best classified models and the remaining ones – note the sudden drop between Study 3 and Study 12. If that sudden drop is clearly due to the loss of the matriculation data (M), while the one witnessed between studies 14 and 6 is due to the loss of the other subgroup of academic data, the curricular data (C).
- Academic data (subgroups CM, Study 13), alone, justify the best study result ($R^2 = 88.4\%$), obtained in *ex-aequo* with Study 8 (CMS).
- As expected, the assertiveness of the model increases consistently in line with the course of students' academic path.
- Although the results of Study 6 (MDSA), the only one not using the attributes of category C, are far behind the others, they clearly confirm that students'

curricular data is the key contributor to the assertive capacity of the model, probably not least because such data is not available at the beginning of the academic path. Still, it is interesting to observe that even at a very early stage of students' academic course, namely in the 1st and 2nd semesters, the average determination coefficient of the model showing the best performance skyrockets from 64.4% to values greater than 80% and 86%, respectively.

In light of the results presented in Table 3 and of the corresponding considerations exposed, it seems pertinent to propose the groups of variables in Study 13 (categories CM) for the predictive model we intend to design. The study achieves the highest determination coefficients using only 2 of the 5 possible categories of variables. Supporting the choice of the group CM (Study 13) as the best study is also the fact that this set of variables also revealed, in additional tests, the lowest Root Mean Square Error (0.966 when the second lowest error was 0.983 and the average 1.105).

The comparative study conducted allowed the exclusion of three irrelevant groups of variables, from a predictive perspective: the demographic, socioeconomic and the access data. In the next stage, we will try to exclude variables presenting a negligible influence in the performance of the predictive model.

4.2 Additional Adjustment of the Model – Selection of Predictors

The performance of any predictive model to be proposed will be all the more valued as the earlier the moment in which it might be applied. In fact, the predictive relevance of a model is based on two crucial aspects: the veracity of its predictions and the degree of anticipation of such predictions. Therefore, it is now important to fine-tune the CM model (which globally showed to be the most assertive) when applied right at the end of the student's 1st semester, as shown in the scheme in Fig. 2. More precisely, we will try to exclude from the set of CM predictive variables (with the subgroup C including only the curricular results of the student's 1st semester) all those which do not contribute positively and significantly to the quality of the model.

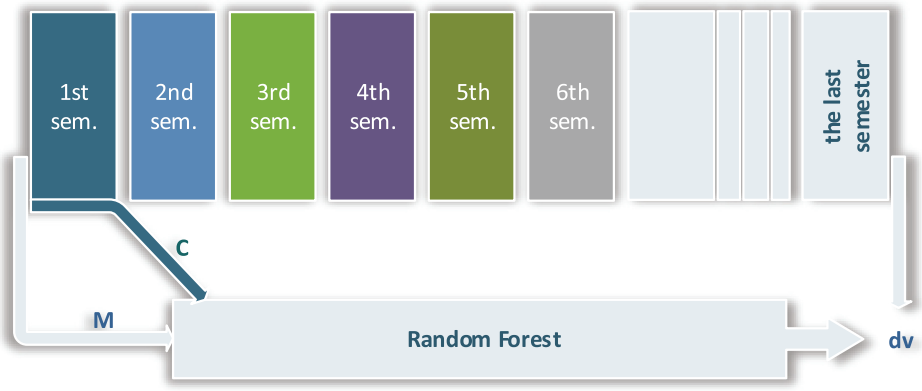

Fig. 2. CM predictive model sustained by data from the 1st academic semester.

Since the CM model did not make use of the socioeconomic and access data, the categories with missing data in a significant number of students, it was possible to use a wider and more comprehensive sample of matriculations containing complete data. Therefore, the size of the data set according to which the CM model will be adjusted increased from 2159 to 4530, the total size of the sample, originating a slight decrease of the model assertiveness in the 1st semester, from $R^2 =$ 80.4 to 79.0. We believe that by using a larger sample of matriculations in this second adjustment of the model, it will present greater generalization capacity.

Before moving on to a more systematised process of fine-tuning of the CM model, the random forest algorithm was run for the data set without variables `n_courses_aprov_s` and `n_courses_disap_s`, for considering them to have a strong correlation with the attributes `ects_aprov_s` and `ects_disaprov_s`, respectively. After confirming the pertinence of excluding these two attributes, the possibility to exclude new variables among those revealing to be less informative, based on random forest ranking of variables in terms of importance, was assessed in successive iterations. The relevant data characterising those several iterations is summarised in Table 4. The data shows that the loss of variables maintained or slightly improved the assertiveness of the model. In a nutshell, we were able to remove 7 out of the 18 attributes from the data set without that affecting negatively the model performance and actually achieving a slight improvement, though of little significance.

Table 4. Removal of variables from the CM data set.

Iter.	Excluded variables		#var	R^2	RMSE
0	`n_courses_aprov_s`	`n_courses_disap_s`	16	79.2	1.334
1	`scholarship_s`	`union_member_s`	14	79.2	1.335
2	`min_s`		13	79.2	1.333
3	`max_s`		12	79.3	1.329
4	`n_assess_disap_s`		11	79.5	1.326

The 11 variables, which together justify the predictive capacity of the model and therefore reveal to be the most determinant in anticipating the academic success of IPB's bachelor degree students were ordered in a descending order of importance (given in brackets) as follows: `ects_disaprov_s` (2.387), `cod_degree` (2.011), `average_s` (1.785), `ects_aprov_s` (1.461), `cred_ects_tx` (1.454), `cod_school` (1.230), `ects_degree` (0.359), `enrol_type` (0.274), `academic_year_s` (0.241), `enrol_year` (0.239), `curricular_year_s` (0.176).

5 Discussion of Results and Conclusions

In this study, the random forest method was used to propose a predictive model of the global academic success of IPB's bachelor degree students at the terminus

of their academic path. Instead of following the commonly adopted procedure of delimiting the prediction to one single specific course, the model was developed from a vast real data set involving records of quite heterogeneous undergraduates from over half a hundred degree courses covering a wide variety of educational fields taught in the five schools composing the institution and where each student is characterized by more than four tens explanatory variables. Such specificity allowed studying the influence of a new curricular factor taken into account for the first time in literature: the type of school. The results obtained allowed concluding that students' success also depends on the school they attend. This conclusion indicates that in order to mitigate retention and academic failure, it might be necessary to adopt differentiated strategies of educational promotion according to each school.

The order of importance provided by the random forest algorithm allowed the identification of the factors contributing to students' success or failure. It enabled the observation that the factors regarding the curricular context of students' academic performance are paramount to the intended prediction, which confirms results previously obtained by [15], who stated that such factors can alone account for academic performance. Note that all 11 attributes which revealed to be significant for the prediction belong, without exception, to the curricular or matriculation categories.

The knowledge obtained allows the identification of students at a higher risk of retention and academic failure, which enables the institutional managers to design more assertive educational or tutorial strategies towards educational efficacy and efficiency.

The kind of approach adopted in the identification of students' characteristics which best account for their success seems to differ from that usually used in works related to the same topic. In the present work, the selection of those characteristics was conducted in two different stages. First, the selection of students' dimensions which best explain their success allowed a first adjustment of the model by eliminating complete groups of variables. Later, a fine-tuned adjustment led to the selection of the attributes which were not excluded in the first stage. This approach enabled us, in addition to reduce the 'plague' of the data dimensionality at an early stage, to exclude completely categories of variables, without losing the predictive capacity of the model. This feature is of particular importance since it contributes to the reduction of the multidisciplinarity of the predictors, thereby lowering some of the complexity of the predictive process, and, more importantly, makes it possible to extend the study to other contexts, where not all categories of variables initially considered in this study are available.

Note however, that a significant part of the results obtained in this study cannot be generalized to the whole context of higher education since they were based on a data sample non-representative of that broader context. This study presented a case study focused on IPB, which for being an institution of the polytechnic higher education subsystem and for being located in an inland region with low population density cannot reach the same heterogeneity of students

as other institutions located in large coastal urban centres. At best, the results presented here may reflect the reality of higher education institutions with similar conditions to those of the IPB such as other polytechnic institutes located in inland regions of the country far from the big urban centres.

Acknowledgments. This work was supported by the Portuguese Foundation for Science and Technology (FCT) under Project UID/EEA/04131/2013. The authors would also like to thank the Polytechnic Institute of Bragança for making available the data analysed in this study.

References

1. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
2. Romero, C., Ventura, S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl. **33**(1), 135–146 (2007)
3. Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(6), 601–618 (2010)
4. Romero, C., Ventura, S.: Data mining in education. Wiley Interdisc. Rev.: Data Min. Knowl. Disc. **3**(1), 12–27 (2013)
5. Baker, R.S.J.D., Yacef, K.: The state of educational data mining in 2009: a review and future visions. JEDM-J. Educ. Data Min. **1**(1), 3–17 (2009)
6. Huebner, R.A.: A survey of educational data-mining research. Res. Higher Educ. J. **19**, 1–13 (2013)
7. Papamitsiou, Z.K., Economides, A.A.: Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. Educ. Technol. Soc. **17**(4), 49–64 (2014)
8. Peña-Ayala, A.: Educational data mining: a survey and a data mining-based analysis of recent works. Expert Syst. Appl. **41**(4), 1432–1462 (2014)
9. Algarni, A.: Data mining in education. Int. J. Adv. Comput. Sci. Appl. **7**, 456–461 (2016)
10. Shahiri, A.M., Husain, W., Rashid, N.A.: A review on predicting student's performance using data mining techniques. Procedia Comput. Sci. **72**, 414–422 (2015)
11. Del Río, C.A., Insuasti, J.A.P.: Predicting academic performance in traditional environments at higher-education institutions using data mining: a review. Ecos de la Academia. **2016**(7), 185–201 (2016)
12. Natek, S., Zwilling, M.: Student data mining solution-knowledge management system related to higher education institutions. Expert Syst. Appl. **41**(14), 6400–6407 (2014)
13. Asif, R., Merceron, A., Ali, S.A., Haider, N.G.: Analyzing undergraduate students' performance using educational data mining. Comput. Educ. **113**, 177–194 (2017)
14. Miguéis, V.L., Freitas, A., Garcia, P.J.V., Silva, A.: Early segmentation of students according to their academic performance: a predictive modelling approach. Decis. Support Syst. **115**, 36–51 (2018)
15. Manhães, L.M.B.: Predição Do Desempenho Acadêmico De Graduandos Utilizando Mineração De Dados Educacionais. Ph.D. thesis (Tese Doutorado), Universidade Federal do Rio de Janeiro (2015)