Decision Support Systems 101 (2017) 1–11

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Early detection of university students with potential difficulties

Anne-Sophie Hoffait*, Michaël Schyns*

HEC Management School, University of Liège, QuantOM, 14 rue Louvrex, 4000 Liège, Belgium

ARTICLE INFO

Article history:
Received 17 May 2016
Received in revised form 4 May 2017
Accepted 4 May 2017
Available online 7 May 2017

Keywords:
Student attrition
Machine learning
Prediction
Classification
Accuracy
Remediation

ABSTRACT

Using data mining methods, this paper presents a new means of identifying freshmen's profiles likely to face major difficulties to complete their first academic year. Academic failure is a relevant issue at a time when post-secondary education is ever more critical to economic success. We aim at early detection of potential failure using student data available at registration, i.e. school records and environmental factors, with a view to timely and efficient remediation and/or study reorientation. We adapt three data mining methods, namely random forest, logistic regression and artificial neural network algorithms. We design algorithms to increase the accuracy of the prediction when some classes are of major interest. These algorithms are context independent and can be used in different fields. Real data pertaining to undergraduates at the University of Liège (Belgium), illustrates our methodology.

1. Motivation and objectives

In an era when manufacturing is no longer the dominant activity, post-secondary education has become crucial to economic success. It is estimated that two-thirds of jobs in future will require a higher education degree. And yet, the 2014 OECD report [1] only records 39% and 11% of graduates among young adults in OECD countries with respectively a tertiary education program of type A (ISCED 5A: university or assimilated) and with one of type B (ISCED 5B: vocational). Despite ever rising registration figures, the drop-out rate generally averages 45% whereas this falls to a mere 10% in specific institutions or fields of study where specific admission criteria apply. As the latter percentage is obtained by restricting admission to the most promising students, this practice cannot be generalized in order to fight the lack of graduated students. On the other hand, where few or no admission requirements apply, the drop-out rate can rise to 70%. In Belgian universities, where admission to most faculties is open to any student holder of a secondary education certificate, where tuition fees are relatively low (about 800EUR annually) and whose reputation attracts a lot of candidates, the average drop rate is about 26.9%. Such figures call for action; reducing them is in order, not only the benefit of the world economy but for the students, the collectivity and the educational institutions as well. Besides damaging self-confidence, failure has a human and a financial cost; one year of study is expensive for students as for the collectivity. As school resources are scanty and diluted among all students, the challenge for educational institutions worldwide is to attract ever more students while getting them through to completion.

The drop-out phenomenon concerns the first year in particular. About 25% of the first generation university students in the French-speaking Community of Belgium drop out after their first year; while some shift to other types of higher education, the majority are in difficulty. Over the past few years, as many as 60% in the French- and German-speaking Community [2] failed their first year. As for the United States, the drop-out rate between the first and the second year reaches 44.8% in two-year institutions and 25.63% in four-year institutions [3]. Worldwide thus, the crucial first year calls for solutions. Remedying the situation, i.e. taking timely positive and efficient decisions, implies understanding the reasons for failure and identifying the students in difficulty as early as possible.

Various reasons can explain failure. Some students fail owing to wrong or inadequate orientation: they do not select the field of study in adequacy with their strengths or the academic or vocational environment may be unsuited to their skills. Others are unprepared to live up to or even understand their program 's demands. Provided these students could be identified and the University could detect their weaknesses in time, relevant and specific academic achievement support could be offered. Most universities, in fact, set up systems meant to help first year students. So far though, it remains difficult to identify critical cases before they run into trouble,

* Corresponding authors.
E-mail addresses: ashoffait@ulg.ac.be (A.-S. Hoffait), M.Schyns@ulg.ac.be (M. Schyns).

http://dx.doi.org/10.1016/j.dss.2017.05.003
0167-9236/

i.e. usually after the first session of examinations, and it is often too late then for reorientation and even adequate help to save the year. This paper aims to detect these students at registration time on the basis on easily available data right then so as to start remediation before the beginning of the academic year or, as the case may be, assist with reorientation. This is a first main originality of our work since most of the previous researches on this subject are usually based on data available only at the end of the first year, or after a session of examinations at the end of the first term, as they have to wait to collect some of the factors [4]. However, the importance of an early detection was already mentioned in some studies [5,6].

Our approach is illustrated by real data provided by the University of Liège. Contrary to other contributions to the issue, we do not focus on any specific fields of study; our extensive database includes a wide range of degrees. We rely on a limited number of indicators of past performances and some environmental factors already identified in the literature. Although we consider a Belgian case, the same indicators could be used for many other countries. In addition, what we propose is a new identification methodology based on data mining methods, which could easily integrate other or additional indicators. Likewise, while our focus is on first year failure at university, our framework can be adapted to other contexts, e.g. vocational studies, distance learning studies, full credit based studies etc.

We use data mining techniques to identify students whose background is the most adverse. We consider three methods simultaneously – namely logistic regression, artificial neural networks and random forest – in comparative perspective. Such methods predicting failure versus success are not new but their level of accuracy is generally too low.

This leads to another objective of this paper and to our second main contribution: we design algorithms to assign observations into a subcategory of special interest with a confidence level predefined by the decision maker. A high level of accuracy can be requested even if it comes at the cost of a smaller set of identified observations. We also show how to deal with the tradeoff between accuracy and size. These algorithms are context independent and can be applied to a broad set of problems. They rely on a dynamic split of the observations into subclasses during the training process. Using subclasses is not a new idea, but we suggest here a dynamic approach where the frontiers of the subclasses are dynamically established based on training data, so as to maximize an accuracy criterion. In the context of the educational case, the process is designed so as to reach a high predefined level of accuracy for the main category under scrutiny, i.e. the failure group.

Distinguishing students with a very high risk of failure from those with a probably lower one has another advantage. Remediation will likely differ in each case as the latter might only need some good advice to succeed, whereas the former might need extra courses. Finally, we analyze results to find out why some students are likely to encounter difficulties. We perform a sensitivity analysis to determine whether they would score better if some characteristics of their profile, including the selected field of study, were modified or improved.

The remainder of this paper is structured as follows. Section 2 displays a brief review of the literature; Section 3 describes methods of classification and their proposed improvement and Section 4 analyzes data pertaining to the University of Liège. Each method is applied on this set and compared. Finally, a sensitivity analysis is performed on some significant factors.

2. Related scientific literature

Researchers set out investigating academic success and drop-out figures years ago, albeit with variations in scope, methods and objective. First, the focus is either on identifying the success factors or, as in our own case, on predicting students' probability of success. Both approaches still share a common interest and the factors identified by the former are particularly relevant as input for the latter. Another question when predicting success is the level of refinement to achieve, i.e. what should be the measure of success. This section starts with a brief review of the literature on success factors and proceeds with success prediction and the measures of success to finally present the common methods towards these goals.

2.1. Success factors

In keeping with our main goal, our review of factors limits itself to papers dealing with factors mainly observable prior to registration. We start with the three categories of factors identified by Arias Ortiz and Dehon [7]. The first contains personal characteristics (gender, nationality, the year of first enrollment and the domain of enrollment); the second one includes high-school path characteristics (the number of years repeated in high school, the type of school, the number of math hours per week and, in the Belgian context, the study or not of Latin or Greek) and the third one integrates socio-economic factors (the household structure, the parents' educational level, occupational activity and income). Some further introduce factors relating to study methodology, student' attitudes [8], perceptions [6], ability and motivation [9] or student's high school grades [10,11]. Inevitably, though, taking these additional factors into account considerably limits the size of the sample because of the difficulty to collect such information and delays or limits the possible remediation actions.

Previous researches do not consider all the factors belonging to the three categories or else restrict their scope. Except for Nandeshwar et al. [12] and Arias Ortiz and Dehon [7], most papers limit themselves to certain academic fields [6,8,13]. Arias Ortiz and Dehon [14] underline the importance of this field factor since students enrolled in Sciences or in Health Sciences in Belgium appear more likely to drop out. These authors also highlight the role of orientation guidance, as shifting to a different field of study after failing a year likely increases their chances of graduation. This is one reason why we perform a sensitivity analysis on this factor. In addition, Suntheim et al. [15] observe different examination cultures across faculties, which might partly explain varying difficulty in succeeding the first year. Integrating factors related to prior schooling is also common. Past performances are expected to be a relevant proxy of future performances. The socio-economic factors are a more controversial issue; while some claim their at least partial relevance to an explanation [7,12,16], others mitigate their significance [9]. Vandamme et al. [6] take the positive view that students can improve their chances of success in spite of such would-be determining pre-registration factors. As well, Bruffaerts et al. [8] observe the growing impact of student attitudes to work upon their success.

2.2. Defining success

Most agree that first year students' results are a relevant measure of their future academic performance, on the grounds that most drop outs occur in the course of that year [17,18]. Arias Ortiz and Dehon [14] find this a reliable measure of academic success since they observe that students failing their freshman year are more likely to drop out. Most still perform the prediction only at the end of the first semester. Some pursue a slightly different objective as they try to predict if a student will graduate or not, such as Suntheim et al. [15]. Márquez-Vera et al. [4] goes a step further by analyzing when the prediction could be the most accurate, depending on the available factors.

Most papers restrict the analysis to two categories: success or failure. Like ours, their emphasis is on failure or drop out [9,14], or both categories. Herzog [18] makes the distinction between students

who stopped/dropped out and those who changed their major. Some try to refine the classification by identifying more classes, e.g. Vandamme et al. [6] define three groups: "the ' low-risk' students, who have a high probability of succeeding; the ' medium-risk' students, who may succeed thanks to the measures taken by the university; and the ' high-risk' students, who have a high probability of failing ". These groups are built *a posteriori*, by examining the link between the scores obtained during the first session of examinations and the success at the end of the year. Despite the addition of subgroups, the level of correct prediction hardly reaches 50% for the failure group. Wolff et al. [19] perform a similar yet more thorough analysis by creating four classes, ranging from the highest to the lowest probability of failing. However, they merely predict the chances of failing for each of these groups. Some differentiate success with and without grade [8] or predict the final grade and examine the probability of completing studies in a specific major [15]. While improving the accuracy of prediction by adding classes is attractive, it has a major drawback : the number of individuals per category that can be used to fit the models decreases to such as point as to become insufficient to obtain reliable prediction by traditional methods.

2.3. Methods

It is quite usual [20] to resort to data mining methods for such problems, whatever the exact goals and sets of factors. When the focus is on analyzing the causes of success or failure (vs. prediction), statistical methods predominate: mainly logit regression e.g. [7,8,15,18] but also e.g. discrete-time survival analysis [9,14]. When the focus, instead, is on prediction, logit regression is still present but other methods like neural networks e.g. [6,17,19], support vector machines e.g. [13,21,22], decision trees e.g. [12,23-25], and random forests [6,17] are also common. Some authors take a step further by applying several methods and comparing their performance e.g. [6,12,13,17]. Typically, all these methods are applied with default parameters. Some, including ourselves, go further still by combining the prediction of several classifiers [22].

While our contribution has its original aspects, it shares some characteristics, especially with the works of Delen [17], Arias Ortiz and Dehon [7] and Vandamme et al. [6]. Our initial selection of factors is based on Arias Ortiz and Dehon [7]. Like them, our scope encompasses first year students enrolled in all fields of study. Our focus, though, is on prediction and not factor analysis. Our context shows that not all socio-economic factors need to be included in the final experiments, for besides being difficult to collect, their contribution does not prove significant. Delen [17] and Vandamme et al. [6] try to identify very early (but during the year) and fairly confidently, the students in potential difficulty. They both compare the results obtained by several data mining techniques. Delen [17] analyzes the case of a public US university for which the average freshmen retention rate is about 80%. The best overall prediction accuracy that he obtains for students who are likely to drop out is 81%. Delen however uses many more variables than us to reach this result, including some that are observable only at the end of the first semester. For instance, his analysis includes the first semester GPA which has been shown to correlate strongly with retention. Vandamme et al. [6] use most of the factors identified in the previous subsection to analyze a Belgian case. Just like ourselves, they suggest adding one intermediate class for uncertainty. Sadly, the rates of correct classification that they obtain are extremely low. They assume that the main reason is due to the limited size of the database. As for us, while using a far larger database but less explanatory variables, we are unable still to significantly increase the quality of the predictions in the two classes.

This is why we have developed new strategies, which we consider our main contribution. Contrary to most authors, who define the additional classes a priori but measure the rate of correct classification a posteriori, we follow a reverse approach where we set the minimal required level of accuracy, and accordingly we dynamically design the classes during the fitting process. As well, we suggest testing several levels of confidence to more closely identify the students in each group. This approach enables us to reach correct classification rates in the failure class above 90%. Finally, we perform a sensitivity analysis in which we assess if a change in the inputs could improve the chances of success. This enables us to test the importance of some factors, in the same spirit as in Thammasiri et al. [21]. We, nevertheless, go slightly further in the analysis so as to suggest potential tracks for remediation. The combination of the detection and the sensitivity analysis in this context is a new contribution.

3. Data mining methods

The link between the key factors and the dependent variable is explored through three classification techniques: logit regression, artificial neural network and random forest. Logit regression and artificial neural network are baseline methods in the education context, as mentioned in Section 2.3. The use of decision trees and random forests is less common even if they have proven their efficiency in terms of accuracy [26] and notably, in the education context [17]. The first step for each of these methods is to fit the parameters for a training data set for which the desired output is already known. We carry out a supervised learning based on historical data. The second step consists in applying the fitted model to new data sets, i.e. students registering. Considering three methods instead of a single one expand usual analysis since it enables us to compare their relative performance [27]. Indeed, any data mining technique could outperform another facing a specific data set or depending on the way the problem is expressed [21].

3.1. Standard methods

3.1.1. Logistic regressions

The logistic regression extends the linear regression to the cases for which the dependent variable is binary (success vs. failure). It aims to predict the probability that an event occurs. After fitting the model, it can be used to predict the probability of success for new individuals but also to identify the most influential factors of success.

3.1.2. Artificial neural networks

An artificial neural network (ANN) is an artificial intelligence technique which is built on analogies with the human neural system. We specifically work with a feed-forward network, designed with only one output neuron, three layers and the logistic activation function [28]. The value of this neuron can be assimilated to the probability to belong to the success class. It usually provides good results with respect to other techniques. The ANN, however, is not always popular since it is extremely difficult to explain and justify a prediction.

3.1.3. Decision trees and random forests

Depending on the answers to a sequence of questions, a decision tree allows to classify an individual into a class, in this case "failure" or "success". A recursive "divide-and-conquer" process is executed to identify the sequence of the most relevant questions. In order to avoid to overfit the model, post-pruning is used [29]. These trees provide extensive information about the decision process and the main key factors (the sequence of questions). The results obtained from a tree are also easier to read than those obtained by other methods.

A random forest is a classifier composed of several trees generated on the basis of independent and identically distributed random vectors [30]. In a classification problem, each tree votes for a class. These votes are combined with the majority voting, i.e. the class attributed to an individual is the most frequent one [31]. The

underlying idea of the random forest is to improve the accuracy of the classification. Hopefully, the forest will offset the error made by some trees. Among the diverse variants, we apply the bagging approach [32], in which samples with replacement are randomly drawn from the data set so that each tree of the forest is trained on one of them.

3.2. Extension to uncertain classes

Our proposal is to design algorithms that provide a high probability of correct classification in the "failure" group at least. Individuals that cannot confidently be associated to the "failure" group or to the "success" group, are set apart in a new "uncertain" class. When possible, this last class is split into the two "uncertain failure" and "uncertain success" subgroups.

As stated in the introduction, such an analysis is needed since the standard data mining approaches do not provide reliable results for the problem at hand. This analysis as well makes it possible to split the population into up to four categories for which the University may decide to take specific (remediation) actions. Our contribution stands in the way of building the different categories. Contrary to Vandamme et al. [6], we do not construct the four categories a priori. Instead, we define the boundaries of these categories during the learning process so as to reach a minimal threshold of confidence.

Our algorithms work separately on the success and failure classes. Improving the accuracy in the failure class does not impact the accuracy in the success group and conversely. Introducing a new class, at the opposite, allows to increase the accuracy in the subgroup of special interest, by moving observations difficult to predict reliably into the "uncertain" class. This last one has typically a lower rate of correct classification. We further motivate this choice in the Results 4.2 section.

3.2.1. Logit regression and artificial neural network

It is straightforward to adapt the logit regression and the artificial neural network for two additional uncertain classes. As the output value computed by the logit regression and by the ANN corresponds to a probability, a value of 1 indicates a better chance of success than a value of e.g. 0.7. Thus we could specify a threshold t_1 above which we are extremely confident that the success prediction is correct and a second threshold t_0 below which failure prediction would be. Any value between these two values corresponds to the "uncertain" class, which can be subdivided into two by defining a third threshold at 0.5.

The main question is how to define adequate values for t_0 and t_1. Note that there is a difference between the individual probabilities stating the chance of success for a specific student, i.e. the value predicted by the model for a specific student, and the probability of correct classification of the model over the whole set of observations belonging to the group of interest, i.e. essentially the failure class in our context. The steps of our algorithm are the following ones:

1. Set the level of confidence C (e.g. $C = 90\%$).
2. Split the learning set into two subsets: 80% for training and 20% for validation.
3. Compute the logit regression over the **training** set.
4. Compute $P(y_i|X_i)$ for each student of the **validation set** and sort the list by increasing order.
5. Set t_0 to the first value of P in the sorted list such that C of the students with $P \leq t_0$ (highest risk of failure) are correctly prognosticated.
   Set t_1 to the first value of P in the reverse ordered list such that C of the students with $P \geq t_1$ (highest probability of success) are correctly prognosticated.
   If no such thresholds exist, reduce C.

Note that we use a validation subset disjoint of the one used for the parametrization of the model to ensure that our thresholds are valid for new unknown observations and not only for those used to compute the model. When we construct a neural network, the output can be interpreted in the same way as the logit regression. The previous algorithm remains valid.

3.2.2. Decision trees and random forests

As for the two previous methods, decision trees allow to compute a probability of success or failure. The probability for being predicted in the class j for an instance x can be estimated as follows :

$$P(j|x) = \frac{N_j(D_x)}{N(D_x)}$$

where D_x is the leaf of the tree reached by x, $N(D_x)$ the total amount of instances reaching that leaf and $N_j(D_x)$ the number of those instances that indeed belong to class j [33]. This approach is easily extended to random forests by taking the average of the probabilities over the different trees for each observation. The algorithm presented above for the logit regression may next be applied.

It is not the only way to construct dynamically the subgroups. We recommend a second approach specific to the decision trees. It is based on the idea that we can change the label of some leaves from "failure" to "uncertain" if the global rate of correct classification is too low. Although our algorithm focuses on failure prediction and does not try to split the "success" class into two categories, the same idea could still be implemented for success.

1. Set the level of confidence C (e.g. $C = 90\%$).
2. Split the learning set into two part: 80% for training and 20% for validation.
3. Construct the tree with the training set.
4. While the global rate of correct classification in the failure class for the validation set is lower than C, select a set of leaves labelled "failure" and change their label to "uncertain" so as to reach the confidence C for the failure group (see below how to perform the selection). This defines a new tree that can be used for prediction into three classes.

Selecting a subset of failure leaves is not trivial. The initial tree could contain many failure leaves. Moreover, the relation between the rate of correct classification in each of the leaves and the global rate of correct classification is complex. For example, changing to "uncertain " the label of the failure leaf with the lowest correct classification rate could have too much impact on the global classification rate and prevent us reaching the level of confidence C; on the other hand, changing to "uncertain " the label of two failure leaves with larger correct classification rates could enable us to obtain exactly the requested level C. The result depends on the number of students routed to each leaf, i.e. the weight of each leaf in the global classification. This becomes a combinatorial problem where any combination of failure leaves should be considered. We have still first considered a heuristic where one leaf was added at a time in the set of leaves to rename. The failure leaf with the lowest correct classification rate at the level of the leaf, i.e. the most uncertain leaf, is selected first. The process is then repeated with the failure leaf corresponding to the second lowest correct classification rate and so on until, as the case may be, we obtain the global confidence C. While this process is fast, there is no guarantee to converge towards the optimal set of failure leaves. Our first numerical experiments confirm that it is extremely difficult to reach C by this approach. Our second approach is based on the fact that computation time is not a strong constraint. The construction of the tree typically has to be done but once a year before registration and we may allot time to the process. We therefore

resort to the exhaustive approach where we build all the possible combinations. We retain the one matching these three criteria:

1. The global rate of correct classification in the failure class is larger or equal to C
2. The global rate of correct classification is as small as possible (close to C)
3. If two sets lead to the same 'optimal' rate, the one leading to the largest number of students identified in the failure class is selected.

The last two criteria ensure that as many students as possible are identified in the failure class (for the required level of confidence). Indeed, defining the classes dynamically has a drawback: greater confidence will typically be obtained at the cost of a smaller set of identified students. A compromise between a large set of students and high confidence has to be achieved by testing different values for C. It is therefore important to get results as close as possible to the requested value of C and not only for larger values. Another advantage to obtain classification for precise values of C is to allow a ranking of the students. The students identified for a large C have a high probability to be identified also for a smaller C. The additional students identified for this smaller value have a lower probability of failure.

While computation time is not a strong constraint, we still cannot afford to apply this process to trees with a very large number of failure leaves. We first limit the size of the trees. A tree node is not expended when the gain in information is not significant enough. Secondly, we set a (large) maximal number of failure leaves to consider in the selection process. Leaves are randomly subsampled when this number is reached.

This new algorithm for the decision trees is fully compatible with a random forest approach. Many trees are built by resampling the learning data set. Each resulting tree makes it possible to sort the data into three classes: high risk of failure, uncertain failure and success. A (new) observation will be associated to the most often encountered class in the forest. This robustifies the approach and even alleviates the effect of suboptimal results which could be obtained by a few trees due to the extra parameter (maximal number of leaves in the optimization process).

4. Case study

4.1. Data set

Our methodology is illustrated by real data pertaining to the University of Liège (ULg). Our focus being on first generation students and first year failure, we set out from an original data set amounting to 11,496 students enrolled there for the first time and spread roughly equally over three academic years, 2011–2012, 2012–2013 and 2013–2014. After disregarding cases of missing data, atypical student status, students who leave during the first months, previous enrollment in other studies (at university level or not) elsewhere and of splitting the first year program into two or more (for administrative reasons, part time study etc), our final sample still counts 6845 students. It is a relatively small data set in comparison with data mining standards but it corresponds to the whole population under scrutiny and cannot be enlarged. This data set is further split into two parts based on a rollover horizon strategy. The first two academic years concerned are used to build the models and the last one to evaluate performances. More generally, the idea is to use historical data corresponding to the last two years in order to make predictions for a new starting academic year.

We rely on a limited number of indicators of past performances and some environmental factors already identified in the literature, especially those identified by Arias Ortiz and Dehon [7]. Few factors are considered not only by choice but also because of the context. Indeed, our aim is to perform an early detection of students in potential difficulty, that is before the beginning of the academic year. Consequently, this restricts the factors that could be collected. This also allows to avoid some privacy issues and to compensate the relatively small number of observations in the data set. While we focus on a Belgian case, the same indicators could be used for many other countries and our methodology remains valid for other study classification based problems. First, some individual characteristics are taken into account: gender, nationality and the field of study, clustered into Human Sciences, Sciences and Health Sciences .[1] Secondly, we have some information about their prior schooling. Looking at their birth date, we compute whether students have registered at the expected age of 17 or later. We also know whether their curriculum included Latin or Greek and which of the three mathematics levels they got. Finally, the only socio-economic factor taken into account is their grant of a scholarship or not. Other socio-economic factors such as their parents' educational level and occupation were also included initially. Yet, these factors are not recorded by default in the registration process and are difficult to obtain. Even when they have been collected for a limited subset of students, they have not improved our results. We see two possible reasons for that fact: first, we think their power of prediction limited, a child's path and profile being more significant than its parents'. Second, increasing the number of factors was done at the cost of reducing of the number of observations while data mining techniques require large training sets for accuracy.

Table 1 summarizes information about the different factors and provides some basic descriptive statistics: the distribution of students into the different modalities and the differential rates of success (DFR). Success rates are similar for each academic year. These rates however vary with the features under the consideration. Girls succeed slightly more often than boys. Belgian students reach a higher success rate than non-Belgians, which is likely due to their non-native command of French and the necessary adaptation to a new environment. The rate of success is not the same for all fields of studies; it is higher in Sciences. Next, the success rates vary with subjects studied in high school. A profile including Latin and Maths turns out to be an advantage.

Thus, the above descriptive statistics show that the factors highlighted here can have an impact on academic success. We further analyze this link by looking at the results of the logistic regression (logit coefficients and odds ratios in Table 1) and look at the estimated coefficients and odds ratios with the same methodology used by Arias Ortiz and Dehon [7]. Practically all factors prove significant, especially the profile including mathematics and the information about age at registration. The Sciences factor is the only one which does not prove significant. By estimating the model, this factor is no longer of great importance, which we could prematurely conclude merely with the DFR. The high proportion of Sciences students having a high level of math can explain the high value of the corresponding DFR. Interestingly, all these initial results confirm those which Arias Ortiz and Dehon [7] derived from for the Free University of Brussels (ULB), as another major Belgian university.

4.2. Results

We now compare the results obtained by applying the four data mining techniques, with and without the extension to uncertain

[1] The Human Sciences include law, political sciences and criminology, philosophy and the arts, psychology, speech therapy and education, business engineering and management sciences, and social sciences. Sciences include architecture, sciences, applied sciences and agronomy. Health sciences include medicine and veterinary medicine.

Table 1
Descriptive statistics of independent variables and results for the logit regression (** Statistically different from zero at 5%. *** Statistically different from zero at 1%.).

		Split %	DFR %	Logit coef.	OR
Individual features					
Academic year					
	2011–2012	33.94	–		
	2012–2013	33.28	0.06		
	2013–2014	32.78	0.82		
Gender					
	Girl	54.36	5.89	0.37***	1.44
Nationality					
	Belgian	95.72	23.84	1.22***	3.4
Studies					
	Human Sciences	49.88	–	–	–
	Sciences	25.27	3.93	−0.11	0.9
	Health Sciences	24.85	−6.74	−0.58***	0.56
Prior schooling					
Late					
	0	76.27	–	–	–
	1	17.24	−22.71	−0.87***	0.42
	2 or more	6.49	−36.83	−1.39***	0.25
Latin/Greek					
	Yes	21.08	18.2	0.55***	1.72
Math					
	Low	7	–	–	–
	Middle	49.1	13.7	0.52**	1.68
	High	43.9	32.3	1.39***	4.02
Scholarship					
	Yes	31.16	−15.18	−0.47***	0.623
Success					
	Success	38.86	–	–	–
Intercept				−2.39***	0.09

classes with a view to identify students likely to encounter difficulties. We do no try to draw conclusions on the impact of some factors, nor potential remediation actions, in this section.

The initial data set is first split into two subsets. A first (larger) set, named learning set, corresponding to the first two academic years is used to estimate the parameters. A second one, corresponding to the most recent year (or available data) and named the test set, is used to measure the quality of the prediction tool. The test set is not used to design the prediction tools. The results on this last set should more faithfully illustrate the capacity of prediction for any prospective student. The learning set itself is subdivided into two sets. The parameters of the standard data mining tools are estimated thanks to the training set. A smaller separate validation set is used in parallel to define the uncertain classes, i.e. to fix the values of t_0 and t_1 (basic dynamic strategy) and to select the failure leaves (algorithm designed for the random forests), as explained above. We used the data mining algorithms of the statistical package R (CRAN) with standard parameters. Following some trials, we decided to design the ANN with six neurons in the hidden layer. The random forest is made of 100 trees with a maximal depth of seven levels. The selection of the best combination of failure leaves is limited to 20 leaves, a dimension that is rarely reached.

Table 2 summarizes the results for the four data mining methods and different levels of confidence. The results are displayed for the validation set, including 921 students, and for the test set, including 2244 students. The first two columns (Std) provide the results for the standard approaches without the uncertain class, i.e. when $C = 0$. This actually leads to a standard classification with only two groups, failure or success. The two next sets of three columns provide the results at a 85% and 90% level of confidence (see Section 4.3 for a discussion about other levels). We show the number of students predicted in each class with respect to the expected class. The rows correspond to the observed/expected classes: failure (F) or success (S). Since the uncertain (UF) class cannot be observed, the corresponding (null) rows are not integrated in the table. The columns correspond to the predicted classes: failure (F), uncertain failure (UF), or success (S). While the success (S) class could have been split into two subgroups by adding an uncertain success class (US), we decided not to include these results in Table 2. The focus of the case study is indeed on identifying students with a high risk of failure (F) and the US class would provide no additional valuable insights. Therefore, class S corresponds here to all the students with a higher probability of success than failure. As a result, Table 2 contains a set of 24 confusion matrices. In each of these 24 boxes, the top left value is the true positive level of prediction for the class of special interest (F). The value just below corresponds to the false positive case. Finally, the "accuracy" lines indicate the percentage of students correctly predicted in each of the predicted categories.

For example, let us look at the logit regression results for the test set. 1037 students in this data set are associated simultaneously to the observed and predicted failure class. 1476 (1037 + 439) students in this data set are predicted as students who will fail, but 439 by error; i.e. 70.3% of correct classification in the failure group. Adding up 1037, 439, 323 and 445 equals 2244, that is the size of the test set. The accuracy for the failure (F) class should always be larger than the value of C for the validation data set since this last set is used to construct the uncertain failure class so as to reach the C level, which is the case. The test set is never used to fit the model but we expect a similar level of accuracy since we hope that the validation set is representative of the population, as for the test set. If differences were observed, we could suspect that at least one of the two data sets is not ideal; either by construction (e.g. a too small or biased representation of each class) or by context (e.g. is the 2013–2014 academic year comparable to the 2011–2012 year?). As a matter of fact, we observe similar results for each set in Table 2, which is very reassuring.

It also comes out that the standard data mining algorithms all provide a correct classification rate around 70% in the failure class. Since the observed rate of failure is 61.14% (see Table 1), this is a poor and unreliable result.

There are significant result differences between the methods. Our new second algorithm for the random forest performs particularly well with an identification of 270 students (12.2%); more than twice the number for the logit regression for a same level of confidence. Moreover, if we accept a slightly lower level of confidence (C = 85%), we identify 455 students (20.1%) as with a higher risk of encountering difficulties. It is also interesting to note that the second algorithm based on relabeling the leaves performs better than the first algorithm based on the rankings, even when looking only at the random forest approach. The leaves algorithm detects more students while maintaining nearly the same level of confidence.

We also observe that we can reach reliability levels up to 90% thanks to our new algorithms. As expected, the drawback is that less students at risk are identified: from 1476 students (65.8%) down to 114 students (5%) for the logistic regression. As already mentioned, greater confidence in the failure prediction unfortunately leads to a smaller set of identified students. We can therefore wonder whether reaching a higher level of confidence is worth it or not. This is an important question and the next section is dedicated to it. A related question is about the level of accuracy in the new UF class. Since our methods assign the students for whom it is more difficult to make a confident prediction to the "uncertain" class, increasing the accuracy in the F class is made at the cost of a low accuracy in the UF class. However, the UF accuracy must be compared with the accuracy obtained without our approach. When we look at the random forest results (leaves) in Table 2, the accuracy decreases from 70.1% (std method) down to 65.9% . In both cases, the accuracy is low and the predictions for these students are not reliable (the observed rate of failure is indeed 61.14%, see Table 1). The split has therefore a limited

Table 2
Prediction results.

		Predicted							
		Std ($C = 0\%$)		$C = 85\%$			$C = 90\%$		
	Observed	F	S	F	UF	S	F	UF	S
Logit									
Validation	F	423	120	121	302	120	44	379	120
	S	176	202	20	156	202	4	172	202
	Accuracy	70.6%	62.7%	85.8%	65.9%	62.7%	91.7%	68.8%	62.7%
Test	F	1037	323	289	748	323	104	933	323
	S	439	445	40	399	445	10	429	445
	Accuracy	70.3%	57.9%	87.8%	65.2%	57.9%	91.2 %	68.5%	57.9%
ANN									
Validation	F	417	126	120	297	126	61	356	126
	S	173	205	20	153	205	6	167	205
	Accuracy	70.7%	61.9%	85.7%	66%	61.9%	91%	68.1%	61.9%
Accuracy	F	1015	345	294	721	345	133	882	345
	S	426	458	47	379	458	16	410	458
	Accuracy	70.4%	57%	86.2%	65.5%	57%	89.3%	68.3%	57%
Random forest (thresholds)									
Validation	F	429	114	149	280	114	83	346	114
	S	184	194	26	158	194	9	175	194
	Accuracy	70%	63%	85.1%	63.9%	63%	90.2%	66.4%	63%
Test	F	1068	292	361	707	292	201	867	292
	S	459	425	64	395	425	22	437	425
	Accuracy	69.9%	59.3%	84.9%	64.2%	59.3%	90.1%	66.5%	59.3%
Random forest (leaves)									
Validation	F	430	113	155	276	112	104	327	112
	S	185	193	22	166	190	9	179	190
	Accuracy	69.9 %	63.1 %	87.6 %	62.4%	62.9 %	92%	64.6%	62.9%
Test	F	1067	293	380	689	291	241	829	290
	S	455	429	75	384	425	29	430	425
	Accuracy	70.1%	59.4 %	83.5%	64.2%	59.4%	89.3%	65.8%	59.4%

Note. Validation set: $n = 921$. Test set: $n = 2244$.

impact on the accuracy in the new UF class and on the decision making process. At least, this poor quality prediction is now limited to a smaller set of students; which is a positive result.

Finally and unsurprisingly, the results for the success class are similar for any values of C since our algorithms were designed and applied to improve the prediction into the failure class and not into the success class. With about 60% of accuracy for a typical observed success rate of 38.9% (see Table 1), this is not exceptional but already nice, especially since the focus is on ' high' risk of failure detection.

By way of illustration, Fig. A2 in Appendix A shows one of the trees belonging to the 90% confidence random forest. This tree is representative of the forest in the sense that it makes the same predictions for 98% of the whole data set. This tree identifies 842 students in the failure class (12.3% of the whole data set) with a correctness rate of 88.8%. Remember that the trees in the forest are not pruned. Some leaves can correspond to very few students and the first levels therefore are more representative. This tree shows that the relations between the attributes and a prediction, i.e. the paths from the root node to a leaf, are not obvious ones. It still confirms that some attributes, like schooling lag, the math level, the field of study and the scholarship play a major role in the prediction.

We have analyzed the results obtained by gathering the predictions of the classifiers. We can more confidently predict if a student is associated to the failure class by two or three of the data mining models instead of just one. However, the maximal percentage of students predicted in the failure class for a same value of C by different methods is obviously the smallest percentage observed independently for each method. The final set could be even smaller but we observe that the students identified by the less performing method, i.e. the logistic regression, are also generally identified by the other methods. The gain in correct classification being small, this approach is discarded.

We conclude this section by a partial answer to a stability question. Can we reproduce these results for other periods of time? We also have access to a similar data set but for the academic year 2008–2009. Using our three algorithms, the same conclusions can largely be drawn. The main difference is that we identify more students in the failure group for a same confidence level. We are able to identify up to 21.2% of students from the test set facing a high risk of failure, with a rate of correct classification of 91%, thanks to the improved random forest approach. Note also that we had access to additional socio-economic factors for a subset of these students. As mentioned, we used this augmented data set to check if the quality of prediction is improved when including these additional variables. This is not the case but the explanation could also lie in a reduced number of observations.

4.3. Tradeoff between the accuracy and the number of identified students

Undoubtedly, a tradeoff exists between the obtained proportion of students classified in the failure group and the level of confidence in this prediction. A greater confidence in the failure prediction unfortunately leads to a smaller set of identified students. Still, there is added information since we can now characterize more precisely two different groups of students. Moreover, our approach provides additional freedom and more guaranties to the decision maker. Our main (second) contribution is indeed to propose an approach that (i) lets the decision maker decide a priori what is the level of confidence that suits him best and, (ii) even more importantly, which provides an optimization process that converges to this requested accuracy. Actually, we also believe that there is no unique answer to the question "what is the suitable tradeoff?". It is case dependent and the decision should be left to the decision maker. An advantage

of our approach is that he can request successively different levels of confidence and take a final decision based on precise information.

We illustrate this approach in the top part of Fig. 1 with the random forest (leaves) and for the test data set. The classification is executed for six different levels of confidence C: 0% (Std), 70%, 75%, 80%, 85%, 90%. Our first claim is that the accuracy observed a posteriori should be close to the value of C requested a priori. It is indeed the case (as already observed for only two levels in Table 2). The observed values are respectively 70.1%, 70.6%, 74.5%, 80.1%, 83.8%, and 89.4%. Our second claim is about the freedom given to the decision maker. The percentage of students who fall in the failure class is provided on the horizontal axis of Fig. 1 for each value of C. In one look, the decision maker can see the impact of a change in the confidence level. He could also do reverse engineering by specifying instead a requested percentage of students in order to determine the corresponding level of confidence. Note also that the students identified with a 90% confidence level are included in the set of students identified with a smaller confidence level. Those identified at $C = 85\%$ are included in the sets for lower confidence levels and so forth. This methodology therefore allows a ranking of the students and a very precise determination of a threshold.

There is another question related to the tradeoff. The number of students identified in the class F for a high level of confidence could be a priori perceived as low and disappointing. However, it is not as bad as the decision maker could believe, as can be illustrated thanks to the lower part of Fig. 1 with the random forest and for the test set. By definition of the problem, the accuracy cannot be measured before the beginning of the year and we must wait for the examination results to compute it. When applying a classical random forest on a fresh cohort of students, the decision maker can therefore only initially assess the quality of the prediction by the probability returned by the forest for each individual. Since a probability of 50% is less reliable than a probability of 100%, the probability attached to each student can be used to define the two subclasses; e.g. if we set a threshold at 90%, then any student in potential difficulty with a probability higher than 90% is attached to the F class and otherwise falls into the UF class. Note that this failure indicator is not exactly the same indicator as the one we use in our methodology (based on the accuracy to belong to the subgroup), but it is the only indication available before the beginning of the academic year. Vandamme et al. [6] use this approach in the education context. In Fig. 1, we use this strategy for a threshold $C = \{0\%, 70\%, 75\%, 80\%, 85\%, 90\%\}$. Note that it is highly probable that no observation has a probability exactly equal to C. Therefore, the observed threshold could be slightly better (50.4%, 70.2%, 75.1%, 80.7%, 85.2%, 90%). As before, the percentage of students who fall in the failure class is provided on the horizontal axis of Fig. 1 for each modality of C. We observe that, when using this other methodology, the number of students identified in the failure class (F) is significantly less than with the approach we propose.

4.4. Sensitivity analysis

In this section, we try to measure the impact of some factors on the prediction. We no longer look directly at the coefficients of the logit regression or at some descriptive statistics as we did in Section 4.1 As a first step towards identifying potential student support actions, we investigate, for the real data set under consideration, what might have happened if a failing student had presented a slightly different profile. Where our models predict success instead of failure when a student shifts to another field of studies, it is interesting to more closely examine this student profile. Education experts talking things over with a student might detect that he might have made a wrong choice and fare better in another field. Likewise, where the models would predict probable success for a candidate who would had taken Latin or mathematics at a higher level in school, looking into such cases could inspire specific remediation actions.

While discussing such possible actions falls outside the scope of the present research, our present aim is to identify a few factors which education experts could fruitfully work on. For, it might be hazardous to draw straight conclusions and design remediation actions on the mere basis of the coefficients or mathematical expressions developed here. Supposing, for instance, that our models predict that a student would have succeeded if he had taken stronger mathematics in school, how is this case to be interpreted? Shall we infer that the solution is to organize extra hours of math for students in difficulty? This might apply if a math knowledge is a prerequisite, as it is usually the case for students enrolled in engineering. If there is a strong case for the underlying rigour of maths to the benefit of any other pursuit or activity, let's not discount the fact that a strong math orientation in high school could be a filter or a test of a student's aptitude or motivation to work. If so, a student's higher probability of success could have less to do with his actual level in math as with his attitude to work. We therefore leave it to specialists define the actions to be taken for remediation.

In this analysis, we have tested three modifications variables of the student profile: reorientation to another field of study, the level in mathematics, or the inclusion of Latin in the school curriculum. Our interest in this subset of factors is twofold. First, these factors are highlighted as significant factors in Section 4.1. Second, contrary to age or gender, it may still be possible to compensate the impact of these three factors by adequate actions. We proceed as follows. Our analysis focuses on students predicted either as "failure " or as "uncertain failure". We modify the value of one of the three inputs under analysis and we look at the new predicted group. We hope to observe a more positive prediction. Some statistics are computed on the number of students that could be helped by this profile modification.

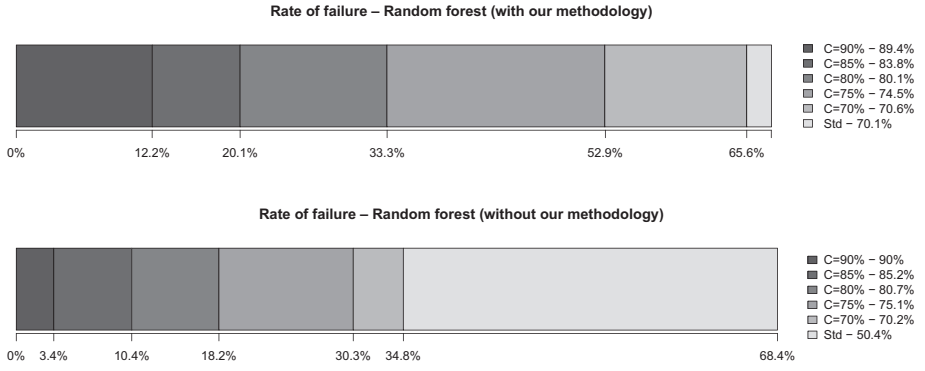

Fig. 1. Percentage of students (test set) identified in the failure group by the random forest with and without our methodology.

Table 3
Descriptive statistics for the three groups.

		All (%)	F (%)	UF (%)	S (%)
Repartition		100	11.82	56.33	31.85
Accuracy		–	89.12	67.43	60.37
Gender	Girl	54.36	51.05	51.79	60.14
Nationality	Belgian	95.72	80.84	96.63	99.63
Studies	Human Sciences	49.88	34.73	61.9	34.22
	Sciences	25.27	18.28	19.37	38.3
	Health Sciences	24.85	46.97	18.72	27.48
Late	0	76.27	15.45	76.66	98.17
	1	17.24	29.91	23.29	1.83
	2 or more	6.49	54.64	0.05	0
Latin	Yes	21.08	11.25	13.54	38.07
Math	Low	7	15.2	9.23	0
	Middle	49.1	58.1	66.03	15.83
	High	43.9	26.7	24.74	84.17
Scholarship	Yes	31.16	44.87	44.24	2.94

The predictions are done thanks to a random forest with a confidence level of 90%; i.e. the best model identified in the previous section. We work on the whole data set (three academic years). 11.82% of the 6845 students are initially predicted in the "failure group" and 56.33% in the "uncertain failure" group. So by gathering both groups, we can carry out the analysis on 68.15% of the students. The distribution of each group of students is provided in Table 3. When we compare the student profiles in each of the groups and in the whole data set, we first observe that the failure group (F) is characterized by a very large number of students being already overage; followed next by the uncertain UF group and nearly none in the success group. This is coherent with the DFR observed in Table 1. Non-Belgians find it more difficult to succeed in their first year. The two failure groups are characterized by most of the students who obtained a scholarship. High math profiles and Latin discriminate the success class. The proportion of Health Sciences students is also above the average in the failure group.

The three attributes under scrutiny are discriminant ones in the previous table. Table 4 indicates how many students are predicted in a more optimistic category if we change the value of one of these three attributes. If we first focus on the students initially associated with a higher risk of failure (F), we see that 312 students over the 809 in this group, i.e. 39%, is now predicted as successful students (F) if exactly one of the three attributes under consideration is changed. If we look at the students initially predicted into the uncertain failure class (UF), 1924 additional students over the 3856, i.e. 50%, could improve their ranking. For the failure class, this result is essentially due to a reorientation for the health sciences students (76% of these students; i.e. 38% of the whole failure group). For the students in the UF group, the level of math is the main reason of prediction improvement (33% of the UF students), followed by Latin on the curriculum (28%) and finally the reorientation (6%, but with a peak to 8% of the Human Sciences students). Note that Table 4 also demonstrates that the relation between one specific attribute and the prediction is not a direct one. It is the combination of attributes which determines the overall profile of a student. It leads to a priori unexpected situations for a handful of students; e.g. decreasing the level of math leads to success for 7 students! We could imagine that these students have acquired other competencies instead of studying advanced math. A closer look into the data reveals that all these students share a similar set of characteristics: they are not over-age, they had Latin in school, they were not on a grant and they were non-Belgians. As shown in Fig. A2, this combination leads to two clearly different paths depending on the level of math. This confirms that a random forest can model more complex situations and can dominate predictions made thanks to a logistic regression tool.

5. Conclusions

We have designed new algorithms relying on data mining techniques to identify the students for which the past context is the most adverse and for which the risk of failure is high. We focus on early detection by limiting the inputs to data that is easily measurable at registration time. This is our first contribution. We consider three methods at the same time, namely logistic regression, artificial neural networks and random forest, and we compare them. We have confirmed that a weakness of these approaches for such a problem is their low rate of correct prediction. We have shown, however, that it is possible to increase the accuracy of the predictions by adding "uncertain" classes. As a second important contribution, we have designed algorithms to assign observations into a subcategory of special interest and with a level of confidence predefined by the decision maker. He can request a high level of accuracy, even if it is at the cost of a reduction of the class size. We have additionally shown that it allows a more precise characterization of all the students with potential difficulties. These algorithms are context independent and can be applied to a broad set of problems. They rely on a dynamic split of the observations into subclasses during the training process, so as to maximize an accuracy criterion.

On a real data set, we are now able to identify with a high rate of confidence (90%) a subset of 12.2% of students facing a very high risk of failure. Moreover, our approach makes it possible to rank the students by levels of risk. Finally, we have performed a "what if" sensitivity analysis to identify more precisely the profile of students facing difficulties and to determine some characteristics on which remediation actions could be built. The final goal is certainly not to confine students in classes, but to identify them in order to give them efficient help.

Acknowledgments

We would like to thank the authorities of the University of Liège and the RADIUS Business Intelligence Unit for their support. The paper, however, only expresses the author's views. This work was partially funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office (grant P7/36).

Table 4
New predictions if one of the attributes is modified.

		Init. F	F to UF	F to S	Init. UF	UF to S	Init. S
Best change of one attribute		809	148	312	3856	1924	2180
Studies	Human Sciences	281	8	2	2387	195	746
	Sciences	148	0	14	747	26	835
	Health Sciences	380	2	288	722	24	599
Latin	No	718	21	5	3334	1079	1350
	Yes	91	48	0	522	0	830
Math	Low	123	0	5	356	123	0
	Middle	470	0	42	2546	1152	345
	High	216	73	7	954	0	1835

Appendix A. Additional material

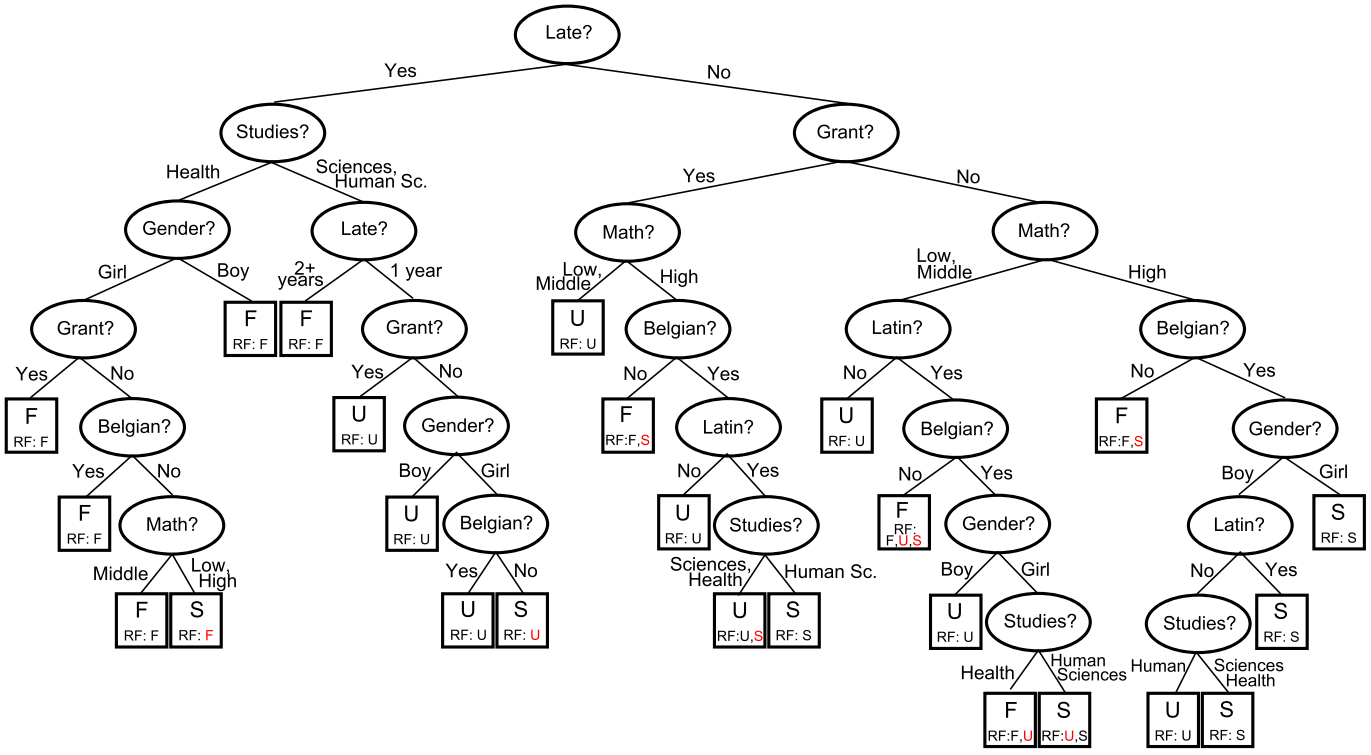

Fig. A2. One representative tree belonging to the random forest. F : Failure, S : Success, U : Uncertain; RF : possible prediction(s) made by the random forest for the same path.

References

[1] OECD, Education at a Glance 2014: OECD Indicators, OECD Publishing. 2014.
[2] Ministère de la Communauté Francaise de Belgique & l'Entreprise des Technologies Nouvelles de l'information et de la Communication, Les indicateurs de l'enseignement, Ministère de la Communauté Francaise de Belgique. 2014.
[3] ACT, National Collegiate Retention and Persistence-to-Degree Rates, 2015.
[4] C. Márquez-Vera, A. Cano, C. Romero, A.Y.M. Noaman, H. Moussa Fardoun, S. Ventura, Early dropout prediction using data mining: a case study with high school students, Expert. Syst. 33 (1) (2016) 107–124.
[5] N. Nistor, K. Neubauer, From participation to dropout: quantitative participation patterns in online university courses, Comput. Educ. 55 (2) (2010) 663–672.
[6] J.P. Vandamme, N. Meskens, J.F. Superby, Predicting academic performance by data mining methods, Educ. Econ. 15 (4) (2007) 405–419.
[7] E. Arias Ortiz, C. Dehon, What are the factors of success at university? A case study in Belgium, CESifo Econ. Stud. 54 (2) (2008) 121–148.
[8] C. Bruffaerts, C. Dehon, B. Guisset, Can schooling and socio-economic level be a millstone to a student's academic success? ECARES Working paper 2011-016, 2011. (21 pages).
[9] G.M. Alarcon, J.M. Edwards, Ability and motivation : assessing individual factors that contributes to university retention, J. Educ. Psychol. 105 (1) (2013) 129–137.
[10] P. Cyrenne, A. Chan, High School Grades and University Performance: A Case Study, Econ. Educ. Rev. 31 (5) (2012) 524–542.
[11] J.L. Demeulemeester, D. Rochat, Impact of individual characteristics and socio-cultural environment on academic success, Int. Adv. Econ. Res. 1 (3) (1995) 278–287.
[12] A. Nandeshwar, T. Menzies, A. Nelson, Learning patterns of university student retention, Expert Syst. Appl. 38 (12) (2011) 14984–14996.
[13] S. Huang, N. Fang, Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models, Comput. Educ. 61 (2013) 133–145.
[14] E. Arias Ortiz, C. Dehon, Roads to success in the Belgian French Community's higher education system: predictors of dropout and degree completion at the Université, Libre de Bruxelles, Res. High. Educ. 54 (6) (2013) 693–723.
[15] K. Danilowicz-Gosele, J. Meya, R. Schwager, K. Suntheim, Determinants of students' success at university, Center for European Governance and Economic Development Research Discussion Papers, 2014. pp. 214.
[16] F. Araque, C. Roldán, A. Salguero, Factors influencing university drop out rates, Comput. Educ. 53 (2009) 563–574.
[17] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decis. Support. Syst. 49 (4) (2010) 498–506.
[18] S. Herzog, Measuring determinants of student return vs. dropout/stopout vs. transfer: a first-to-second year analysis of new freshmen, Res. High. Educ. 46 (8) (2005) 883–928.
[19] A. Wolff, Z. Zdrahal, D. Herrmannova, P. Knoth, Predicting student performance from combined data sources, in: A. Penã-Ayala (Ed.), Studies in Computational Intelligence, 524, 2014. pp. 175–202.
[20] A. Peña-Ayala, Educational data mining: a survey and a data mining-based analysis of recent works, Expert Syst. Appl. 41 (4) (2014) 1432–1462.
[21] D. Thammasiri, D. Delen, P. Meesad, N. Kasap, A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition, Expert Syst. Appl. 41 (2) (2014) 321–330.
[22] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, V. Loumos, Dropout prediction in e-learning courses through the combination of machine learning techniques, Comput. Educ. 53 (3) (2009) 950–965.
[23] D. Delen, Predicting student attrition with data mining methods, J. Coll. Stud. Retent. 13 (1) (2011) 17–35.
[24] P. Schumacher, A. Olinsky, J. Quinn, R. Smith, A comparison of logistic regression, neural networks and classification trees predicting success of actuarial students, J. Educ. Bus. 85 (5) (2010) 258–263.
[25] E.N. Maltz, K.E. Murphy, M.L. Hand, Decision support for university enrollment management: implementation and experience, Decis. Support. Syst. 44 (1) (2007) 106–123.
[26] L. Rokach, O. Maimon, Data Mining with Decision Trees. Theory and Applications, Series in Machine Perception and Artificial Intelligence, World Scientific Publishing Company, Incorporated. 2011,
[27] D.L. Olson, D. Delen, Y. Meng, Comparative analysis of data mining methods for bankruptcy prediction, Decis. Support. Syst. 52 (2) (2012) 464–473.
[28] M. Titterington, Neural Networks, Wiley Interdiscip. Rev. Comput. Stat. 2 (1) (2010) 1–8.
[29] L. Breiman, J.H. Friedman, Classification and Regression Trees, Wadsworth and Brook. 1984.
[30] L. Breiman, Random Forests, Mach. Learn. 45 (1) (2001) 5–32.
[31] P.N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson, Harlow, 2006.
[32] L. Breiman, Bagging Predictors, Mach. Learn. 24 (2) (1996) 123–140.
[33] D.D. Margineantu, Class probability estimation and cost-sensitive classification decisions, Mach. Learn.: ECML 2002 2430 (2002) 270–281.

Anne-Sophie Hoffait is currently a PhD candidate with HEC-Liège, the management school of the University of Liège, Belgium. She obtained her MSc degree in Mathematics from the University of Liège. Her research focuses mainly on machine learning for business decision making.

Dr. Michaël Schyns is a full Professor in Management Information Systems at HEC-Liège, the Management School of the University of Liège (Belgium).

His research interests are in Business Analytics and Combinatorial Optimization. He was elected President of the Belgian Operational Research Society for 2017–2018.