Decision Support Systems 115 (2018) 36–51

Contents lists available at ScienceDirect

Decision Support Systems

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

Early segmentation of students according to their academic performance: A predictive modelling approach

V.L. Miguéis[a,*], Ana Freitas[b], Paulo J.V. Garcia[b], André Silva[b]

[a] Faculdade de Engenharia da Universidade do Porto, INESC TEC, Rua Dr. Roberto Frias, Porto 4200-465, Portugal
[b] Faculdade de Engenharia da Universidade do Porto, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

ARTICLE INFO

Keywords:
Educational data mining
Predictive modelling
Data mining
Academic performance
Engineering education

ABSTRACT

The early classification of university students according to their potential academic performance can be a useful strategy to mitigate failure, to promote the achievement of better results and to better manage resources in higher education institutions. This paper proposes a two-stage model, supported by data mining techniques, that uses the information available at the end of the first year of students' academic career (path) to predict their overall academic performance. Unlike most literature on educational data mining, academic success is inferred from both the average grade achieved and the time taken to conclude the degree. Furthermore, this study proposes to segment students based on the dichotomy between the evidence of failure or high performance at the beginning of the degree program, and the students' performance levels predicted by the model. A data set of 2459 students, spanning the years from 2003 to 2015, from a European Engineering School of a public research University, is used to validate the proposed methodology. The empirical results demonstrate the ability of the proposed model to predict the students' performance level with an accuracy above 95%, in an early stage of the students' academic path. It is found that random forests are superior to the other classification techniques that were considered (decision trees, support vector machines, naive Bayes, bagged trees and boosted trees). Together with the prediction model, the suggested segmentation framework represents a useful tool to delineate the optimum strategies to apply, in order to promote higher performance levels and mitigate academic failure, overall increasing the quality of the academic experience provided by a higher education institution.

1. Introduction

Considering that one of the Europe 2020's targets states that at least 40% of the population aged 30–34 should have completed a tertiary education by 2020 [23], and that one of the USA's goals for 2020 is to lead the world in college graduates [3], higher education institutions are faced with the challenge of, alongside attracting more students, dealing effectively with their very different academic performances. Within this challenging scenario, institutions have to timely devise strategies to promote academic success and enhance the academic experience of students with different academic performance levels.

To achieve this, higher education institutions are now becoming aware of the potential of studying educational data to improve the quality of their managerial decisions [34,19,37]. They are making efforts and investing in creating information systems to collect education-related data, and studying it using data mining techniques. The purpose is to extract meaningful and operational information from those large educational databases [28], which have the potential to provide a better understanding of students' behavioural patterns.

Despite the promising potential of data analysis supported by data mining, most higher education institutions have not been able to analyse this data and transform it into valuable information. In fact, in most cases, only conventional methods supported by statistics have been applied to the data. Therefore, the use of these new generation techniques is clearly in the agenda of higher education institutions, in order to support the development of educational strategies [25].

Among the tasks where educational data plays an important role and where the literature has already made some progress, we can find the prediction of academic performance. A few institutions are now aware that the early inference of students potential academic performance may enable them to foster higher levels of academic achievement. This may result in the design of differentiated actions targeting different groups of students according to their potential and may also result in a more efficient allocation of the institutions' resources.

In this context, this study aims at supporting an European Engineering School in promoting the academic potential of each

[*] Corresponding author.
E-mail addresses: vera.migueis@fe.up.pt (V.L. Miguéis), anafreitas@fe.up.pt (A. Freitas), pgarcia@fe.up.pt (P.J.V. Garcia), ei10085@fe.up.pt (A. Silva).

https://doi.org/10.1016/j.dss.2018.09.001
Received 29 July 2017; Received in revised form 1 August 2018; Accepted 3 September 2018
Available online 18 September 2018
0167-9236/

student and experience through data mining models. For this purpose, this study's main goal is to early classify students into segments, which are not only based on the students average grade but also on the time taken to conclude the degree. It can be argued on higher education system efficiency and on the academic experience quality grounds that the average grade is incomplete and therefore the identification of those more likely to take a long time to graduate allows target intervention programs to act where they are needed most. Particularly, this study addresses the following questions:

1. Is it possible to identify, in an early stage of the students' academic path, their future academic success groups?
2. In the specific context of study, which data mining technique, among random forests, decision trees, support vector machines, naive Bayes, bagged trees and boosted trees, performs best?
3. What are the dimensions that mainly determine the propensity of students to achieve a certain academic performance level?

This study adds value to the literature in several dimensions. First, and in what regards the methodology, we propose a two-stage approach, combining discretization and classification methods. Initially, using a discretization algorithm, the students' performance is categorized into five classes, corresponding to different levels of academic performance. Then, we propose a model to early predict the performance of engineering students at the end of the academic degree or at an advanced stage of their academic career. The students' performance is based on a new performance metric which combines the mean of the grades obtained and the number of enrollments in the courses to achieve that mean.

Several classification techniques are applied, namely random forests, bagged trees and boosted trees, to predict the overall level of academic performance of the students. The application of ensemble methods, such as these previously mentioned, to the educational data mining field is still incipient, although their predictive performance is generally high.

Third, we assess the importance of student-related variables to the prediction model in order to provide decision makers some information regarding the factors that impact students performance, both in terms of GPA and time to degree completion.

Finally, this study contributes to the literature, by proposing an approach to segment students based on both the predicted performance at the end of the academic degree, and the performance observed at the end of the first year. This segmentation approach may be a basis for the differentiation of the actions developed by the institutions to: not only promote higher levels of academic success, but to enhance the quality of the students educational experience. For example, when designing a program to promote higher success, the institution may be interested in targeting a student belonging to the lowest performance group according to the prediction model, and that, simultaneously, already demonstrates signs of poor performance at the beginning of the degree program. At the other extreme, when designing a reward program, an institution may be interested in targeting students who are top performers according to the prediction model, and that are already top performers at the end of the first academic year.

The present study differs from other works on academic performance, as it is based on a high number of students, i.e., 2459 students, corresponding to five cohorts of students who enrolled in a European engineering and technology school during five academic years (2003 to 2007, followed up to 2015); most of the studies in the literature consider small sets of students to validate the proposed models.

The paper is structured as follows. The following section presents the related studies, in order to emphasize the contributions of the current study. Section 3 introduces the methods and data, the variables included in the proposed model, and the performance evaluation criteria. Section 4 addresses the results and the discussion. Section 5 highlights the conclusions and Section 6 the limitations and ideas for future research.

2. Related studies

As huge amounts of data are being made available by the institutional information systems, data mining techniques emerge as natural tools to tackle the above questions. The use of data mining in this context is not new (see Table 1). The progress in the educational data mining research field can be followed in several reviews (e.g. [58,57,69,68]). These reviews provide many examples of the application of the different data mining techniques to support a variety of aspects related to education. These aspects include, for example, students' dropout prediction (e.g. [19,46, 75]), the development of recommendation systems (e.g. [1,22]), and students' performance prediction (e.g. [35,45]). Regarding dropout prediction, for example Márquez-Vera et al. [46] conduct several experiments using a dataset of 419 students to predict dropout at different steps in a course, and select the best predictors of dropout. Bydzovska [15] develops a recommendation model based on the students' skills, knowledge, interests and free timetable time-slots to support students in their choice of selective and optional courses; this model is validated using data from 1444 students.

The prediction of students' performance is one of the most popular and useful applications of educational data mining. This consists of estimating the unknown value of students' performance, score or mark [70]. This is a challenging problem to solve, due to the large number of circumstances that can impact students' performance, such as socio-economic status, previous scholar experience, interactions between colleagues, demographic characteristics, psychological profile and cultural background [4,64]. Tinto [76] introduces a predominant theoretical framework regarding academic success. This popular framework considers academic success as a socio-psychological connection between the characteristics of the student enrolling in the university and the experience at that institution. According to this model, the commitment to the institution and the commitment to the goal of study completion are highly connected to the degree of integration.

Considering the aforementioned factors, the educational data mining studies that focus on students' performance have been using a high number of attributes to characterize students and their environments. Bordea et al. [10] review the attributes in predicting students' performance in a course. Internal assessment attributes, such as assignment marks, quizzes, lab work, class tests and attendance, are frequently used among the researchers to predict students' performance [55,56]. Bordea et al. [10] also stress students' demographic attributes and external assessment attributes as relevant features. Demographic attributes include gender, age, family background, and disability [51,16,36]. External assessments correspond to marks obtained in the final exam for a particular subject. Furthermore, attributes related to high school background [29,72], social interaction networks [61,67] and extra-curricular activities [51,36,47] are frequently considered. There are also several studies that utilized psychometric factors to predict students' performance [26,50]. A psychometric factor is identified as a student interest, study behaviour, engagement time, and family support.

Regarding the concept of students' performance, a branch of the literature is dedicated to exploring the success in a specific course [35,45,18]. In this case, several studies define academic success according to the collected average point grade (e.g. [35]), while other studies only consider whether a student failed or passed in that specific course (e.g. [45,18,44]). There is another, smaller branch of the literature, that explores academic performance at the degree level. In this case, some studies aim to predict whether a student will get a degree [2], while others aim to predict a student's final grade (e.g. [28,40]). Several other studies focus on determining the students' academic performance at the end of the first academic year [34,26,77].

In order to analyse the gathered educational data, several data mining techniques can be applied. Classification and regression are those mostly applied when handling a problem of performance

Table 1
Studies addressing students' academic performance.

Study	Main objective	Techniques*	Performance focus
Huang and Fang [35]	To predict student's scores on three dynamics mid-term exams	LinR, NN, SVM	Performance in exams
Marbouti et al. [45]	To identify at-risk students in a course that used standards-based grading	LogR, NN, SVM, DT, NB, KNN	Success in course
Costa et al. [18]	To predict students likely to fail two courses (one performed on campus and another in a distance education format) at an early enough stage	NN, SVM, DT, NB	Success in course
Gray et al. [26]	To identify college students at risk of failing in the first year of study	LogR, NN, SVM, DT, NB, KNN	First year performance
Macfadyen and Dawson [44]	To identify which student online activities accurately predict academic achievement	LogR	Success in course
Guruler et al. [28]	To categorize students as either successful or unsuccessful and determine profiles of students whose GPA is equal to 2.0 (which is the minimum GPA required for graduation) or greater and of those students whose GPA is equal to 3.0 or greater (honor degree)	DT	Degree level performance
Laugerman et al. [40]	To determine what academic integration characteristics contribute to their success in engineering using post-hoc graduation data	BR	Degree level performance
Hoffait and Schyns [34]	To identify freshmens' profiles likely to face major difficulties to complete their first academic year	LogR, NN, RF	First year performance
Romero and Ventura [70]	To predict the marks that university students will obtain in the final exam of a course	DT, NN, RI	Success in course
Palmer [54]	To predict academic performance of engineering students enrolled in a second-year class	LogR	Degree level performance
Romero et al. [66]	To predict students' final marks based on their participation in forums	LogR, NN, RF, NB, bayesNet, SMO	Success in course
Mishra et al. [50]	To predict the third semester performance of MCA students	DT, RT	3rd Semester performance
Natek and Zwilling [51]	To predict the success rate of students enrolled in their courses	DT	Success in courses
Arsad et al. [5]	To predict the academic performance of Electrical Degree students	NN	8th Semester performance
Strecht et al. [74]	To predict approval/failure in a course and to predict its grade	LinR, NN, SVM, DT, NB, KNN, adaBoost	Success in course
Vandamme et al. [77]	To predict the first year performance of students	DA, NN, RF, DT	First year performance
Aluko et al. [2]	To predict academic success of architecture students based on information provided in prior academic performance	DA, KNN	Degree level performance
Current study	To predict performance levels in the end of the degree (or at an advanced stage of the academic career), in the end of the first academic year	RF, DT, SVM, NB, BagT, BoosT	Degree level performance

*LinR: Linear Regression, NN: Neural Networks, SVM: Support Vector Machines, LogR: Logistic Regression, DT: Decision Trees, NB: Naive Bayes, KNN: k-nearest neighbor, RI: Rule Induction, BR: Boosted regression, SMO: Sequential Minimal Optimization, RT: Random Trees, BagT: Bagging Tress, BoosT: Boosting Trees.

prediction. Classification is used when the predicted variable is a categorical value, while regression is used when the predicted variable is a continuous variable. Among the classification and regression techniques used for classification and regression purposes, decision trees, artificial neural networks, naive Bayes, k-nearest neighbour and support vector machines are the most frequently used [10]. Decision trees are used, for example, by Mishra et al. [50] to predict the third semester performance of a group of 250 master program students, and by Natek and Zwilling [51] to explore the use of small student data sets (i.e., two samples of 32 students and one of 42 students) to predict students' performance in informatics courses. Regarding Neural Networks, Arsad et al. [5] propose a model to predict the achievement of 505 students in the eighth semester, using data collected in the first semester. Naive Bayes is used by Marbouti et al. [45] to identify at-risk students in a particular course, using a sample of approximately 1600 students. Support Vector Machines is also applied by Gray et al. [26] to identify college students at risk of failing in the first year of their studies; this study uses a dataset of 1074 students. Strecht et al. [74] also apply Support Vector Machines to predict students' success in a set of 391 courses.

Table 1 summarizes the educational mining studies that address students' academic performance, and are mentioned above. This table mainly focuses on the type of performance evaluated as well as the techniques employed. An extended version of Table 1 is available in the Appendix A.1. This table also highlights the number of observations (in most cases, the number of students) used in the study and provides further details on the dependent variable used in the study.

The analysis of the literature reveals that there are opportunities for development in several domains. First, a considerable number of the studies address prediction performance in a specific course, while those addressing overall performance are mainly based on the GPA obtained. Thus, there is room for exploration of different performance measures, as the average grade does not reveal much about the academic career of students, namely their failures in courses and the time taken to obtain the degree. Second, most of the studies available in the literature do not explore the use of state of the art classification techniques such as the ensemble techniques, which have revealed a huge potential in other contexts. In addition, most of the studies in the literature consider small sets of data to validate the proposed models. Finally, in general, the literature does not discuss the process that follows the classification of the students in terms of academic performance. For example, if a student is predicted to have poor performance in the future but, until the moment of the application of the prediction model is performing well, the level of actionability of the student is conditioned. Consequently this student should be managed in a different way when comparing to another student that in the moment of the prediction is already having poor performance.

3. Methods and data

3.1. Proposed method

This study aims to perform an early segmentation of students according to their academic performance. For this purpose, we propose to use a two-stage approach, illustrated in Fig. 1.

Firstly, students are grouped into sets, considering their overall academic performance indicator (defined in Section 3.4). We use a binning algorithm (equal-width binning) to establish five levels of

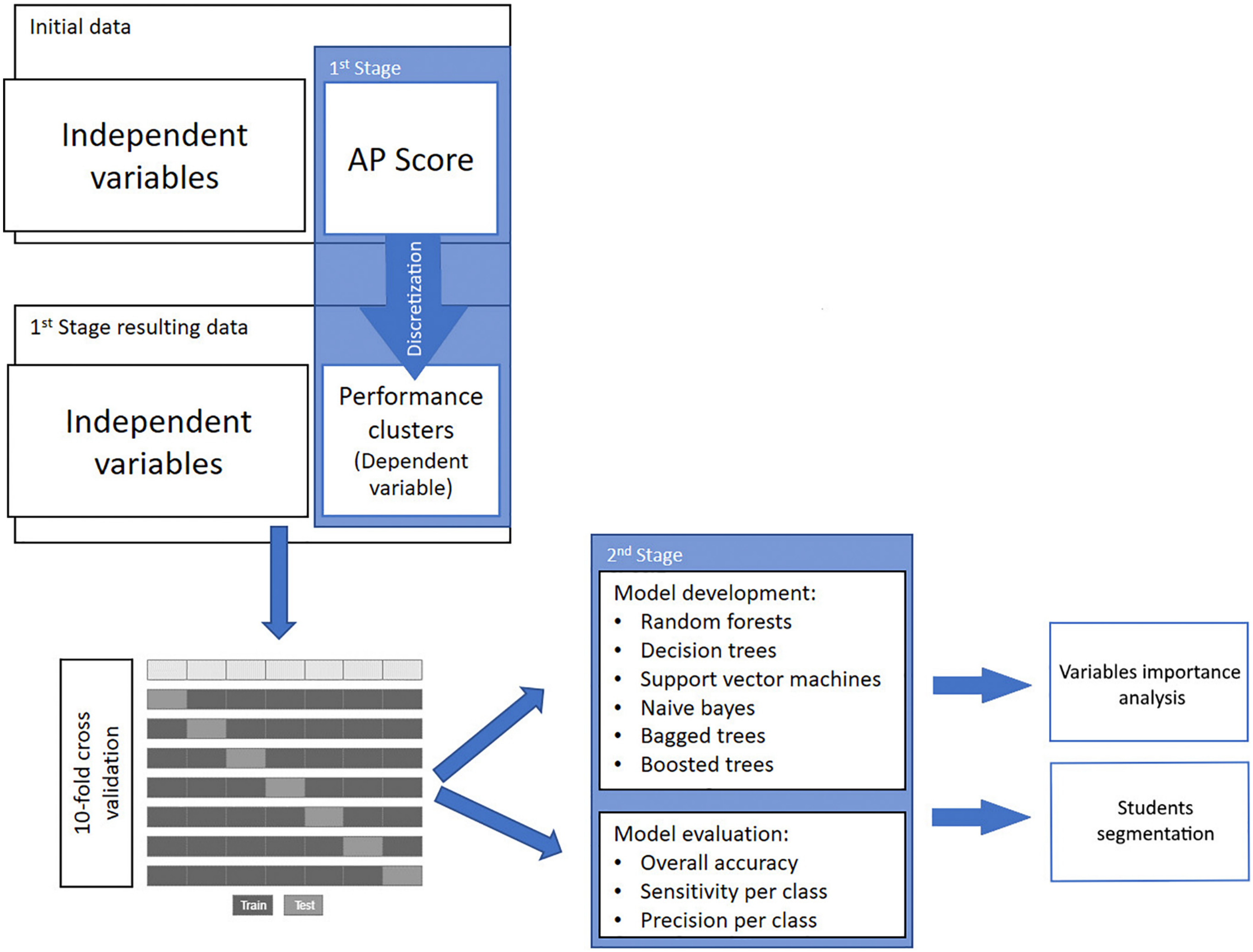

Fig. 1. Proposed model.

academic performance, i.e., A, B, C, D and E. [1] The identification of five academic performance levels is mainly related to managerial questions. This is the maximum number of groups considered manageable by the institution and that, at the same time, enables to discriminate the students with reasonable detail.

Considering the numeric overall academic performance indicator, the equal-width binning algorithm creates a predefined number of groups (in this case five), by dividing the range of AP score into intervals of equal size. Note that there are alternative methods to discretize the numeric academic performance indicator [21].

It is important to note that our goal is not to determine the exact performance value of a student, but rather to be able to segment them, as determined by grouping the performances of the students' peers. The performance level, obtained through the binning procedure, corresponds to the dependent variable of the prediction model developed in the second stage. This means that this is the variable we intend to predict in the end of the first year of the students' academic career.

Secondly, we construct a data mining prediction model that uses the data available at the end of the first year as independent variables, and uses, as a dependent variable, the categorical variable reflecting the level of academic performance at the end of the degree assigned to each student, as described above. Thus, this model constitutes a multi-class classification problem.

Following the literature, we propose to include, in the prediction model, independent variables related to the high school background, psychometric factors, social factors and demographics. We also propose to incorporate the results of internal and external assessments of the courses that are part of the first semester of the academic program.

Besides the proposed two-stage approach, this paper aims to segment students according to their first year observed performance, and their expected performance. Thus, we propose to use a matrix that confronts the students' predicted performance levels with the different level of academic success already observed in the first year of their studies. The indicator used to group students according to their academic achievement in the first year was the same used to group students at the end of the degree (defined in Section 3.4). We believe that each student segment leads to different potential actions that can be used to promote higher academic success, or to reward students with proven results.

3.2. Data

This study uses secondary data from an engineering and technology school, belonging to a European public research University[2]. The school

[1] Note that this procedure is not the same used for the ECTS grade conversion, namely in the context of student mobility, such as in Erasmus+, nor standard GPA.

[2] Faculdade de Engenharia da Universidade do Porto, Portugal.

has approximately 7000 students, 500 academic staff and 300 researchers, and offers undergraduate and graduate (masters, doctorate) programs in several fields, such as Civil Engineering and Mechanical Engineering.

The data refers to the student information for those enrolled between 2003 and 2007, i.e., 2459 students. It encompasses all academic information obtained until either the conclusion of their degree, or until the academic year 2014/2015. If a student concludes the studies before the academic year 2014/2015, we only collect data until the conclusion, while if a student concludes the studies in the academic year 2014/2015 or after, the last data available refers to 2014/2015. In this study we are not considering the students who have dropped out.

Despite the relevance of all attribute categories highlighted in Section 2, the data provided by the institution used as case study only include socio-demographic information about the student and student socio-economic status, information regarding the high school background, information about the enrolment process and information regarding external assessments. A detailed description of the attributes used in this empirical study is presented in Table A2 in the Appendix A.2.

Regarding socio-demographic information, students are characterized by their gender and marital status. In terms of socio-economic status, we try to infer it by considering their parents' jobs and educational levels. High school background is qualified by the type of school attended (public or private), enrolment average grade, enrolment exams grade and high school final grade. Regarding the enrolment process, students are characterized according to the year of entry into university, the program they enrol in, the enrolment stage and the enrolment option. In the enrolment process, students are required to rank five degree programs according to their preference, in order to allow students with better performance in high school to have priority in selecting a degree program. Furthermore, this process is composed of multiple stages, and the students who are not able to enter in the first stage are conditioned to the number of vacancies that are left in the subsequent stage.

Regarding external assessments, it was only possible to access information in an aggregated level. Thus, students are characterized by their grade at the end of the first and second semesters of the first academic year, as well as by the number of European Credit Transfer System (ECTS) credits referring to the courses concluded in each of these two semesters. Moreover, we could also collect the grade of the student at the end of the academic path, or by end of the period of analysis, i.e., 2015, and the total number of ECTS credits the student enrolled in during this period. These two latter metrics are used to infer the dependent variable of the model.

Fig. 2 illustrates the independent and dependent periods considered in this study, and the period of time in which each variable is collected.

3.3. *Classification techniques*

Six data mining classification techniques are used to predict the performance level each student will achieve: random forests [12], decision trees [62], support vector machines [33], naive Bayes [42], bagged trees [11] and adaptive boosting trees [24]. We use random forests because, despite its general ability to provide very good results (typical of ensemble methods), this technique is underutilized in the educational mining context (see Table 1). Decision trees, support vector machines and Naive Bayes are also used, due to their popularity in the literature, and due to the fact that nowadays the trade-off between their performance and training effort is reasonable [7]. In fact, these techniques do not have any hyperparameter, or have only a few. Bagged trees and adaptive boosting trees are used due to their ensemble nature, usually resulting in higher performances than individual techniques.

Ensemble methods [20], such as random forests, bagged trees and adaptive boosting trees are techniques that combine several models into one final predictive model, generally in order to improve the predictive performance. The motivation for the adoption of ensemble techniques relies in the fact that a set of models with similar training performances may have different generalization abilities. Furthermore, the combination of outputs of several models reduces the risk of selecting a poorly performing model.

In the next sections we briefly describe each of the techniques considered in this study. Please note that the models were run in RapidMiner software and that we have usually used the parameter values defined by default by this software [63].

3.3.1. *Naive Bayes*

Naive Bayes [42] is part of a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the explanatory variables. Bayes' theorem describes the probability of an event, based on conditions that might be relevant to the event. Therefore, given an object to be classified, characterized by several explanatory variables, Naive Bayes assigns to this object probabilities for each of the possible classes.

3.3.2. *Support Vector Machines*

A Support Vector Machine (SVM) [33] is a technique that, given a set of objects belonging to one of two categories, constructs a hyperplane in a high dimensional space that separates those categories. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-data object of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier. SVM were originally introduced to handle binary classification problems. Therefore, several methods were proposed to extend binary SVM to solve multi-classification problems. One popular approach for doing so is to reduce the single multi-classification problem into multiple binary problems.

Support vector machines usually imply the definition of metaparameters such as the kernel and the complexity constant. In this particular case we used a linear kernel function with a complexity constant equal to one.

3.3.3. *Decision trees*

Decision trees [62] are classification techniques based on the concept of divide and conquer. This means that the initial set of data is divided progressively in smaller subsets. These subdivisions are based on the values of an explanatory variable chosen according to an attribute-selection criterion, i.e. criterion which identifies the attribute that "best" separates a given data set of objects into individual classes. For each subset a child node is created and the subset data is included in it. The process is then subsequently repeated on the data of the child nodes, until a termination criterion is satisfied. A decision tree has a tree structure, where each node is either a leaf, which indicates the value of the class, or a decision node, which specifies some test to be carried out on a single explanatory variable. Each outgoing branch represents an outcome of the test.

3.3.4. *Random forests*

Random forests [12] are part of the ensemble techniques group, since they combine multiple decision trees in order to outperform the performance of each individual decision tree. Usually this technique enables to overcome problems related to overfitting and noise in data, that are difficult to overcome by a single tree. Each decision tree used by the random forests algorithm is generated based on different training sets which are drawn independently with replacement from the original training set. Moreover, each decision tree is generated by considering a random sample of attributes. Each decision tree gives a classification for each object, called "vote" for that class. The random forest assigns to each object the class having a higher number of votes (over all trees in the forest).

In this case we have adopted a number of trees equal to 100 and the number of random features to pick at each node split was defined as five.

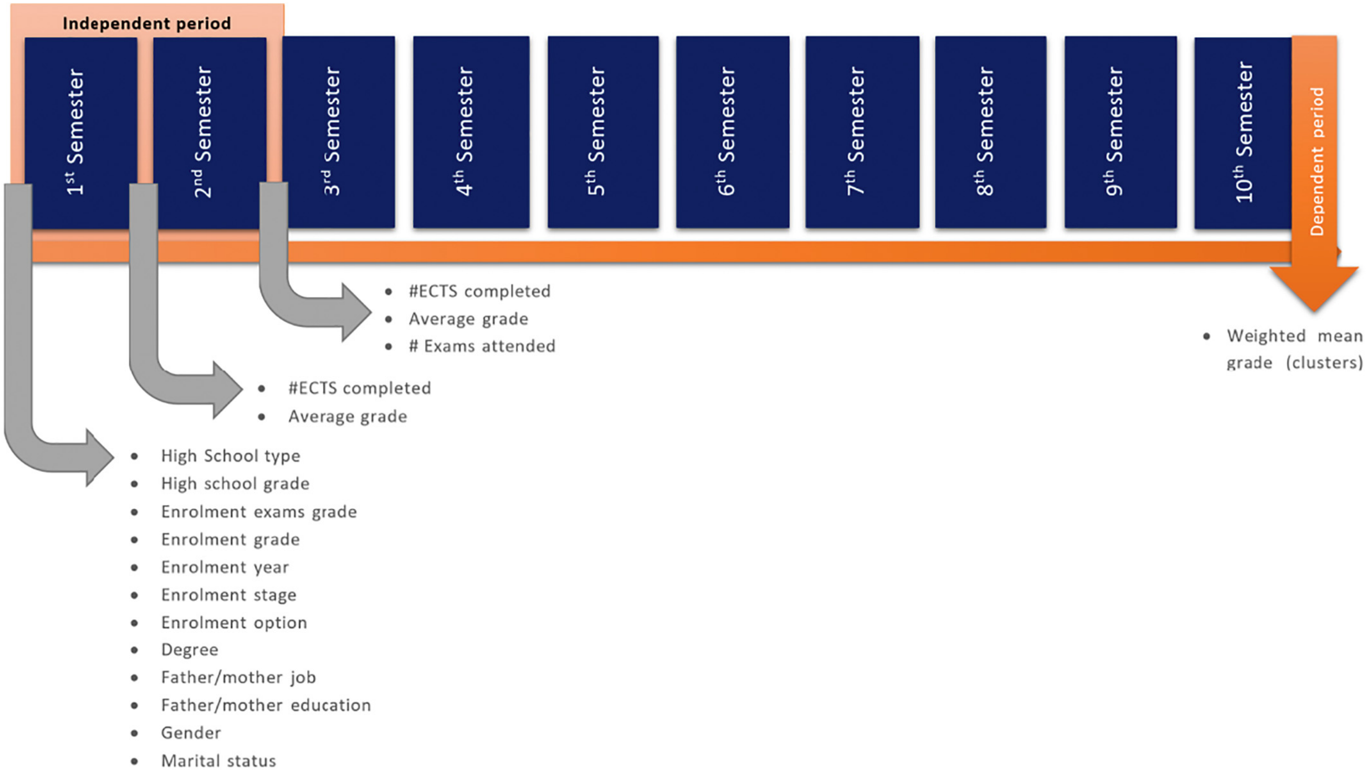

Fig. 2. Prediction model representation.

3.3.5. *Bagged trees*

Bagged trees [11] are also part of ensemble techniques group. Several decision trees are constructed based on different datasets resulting from a resampling procedure with replacement from the original datasets. This bootstraping procedure results in the exclusion of around 1/3 of the records in each sample. Usually, the classification of an object is also based on the majority vote.

In this case we have also adopted a number of trees equal to 100.

3.3.6. *Adaptive boosting trees*

Adaptive boosting [24] is another ensemble technique, that mainly differs from bagging trees due to the fact of not using random subsamples of the data to construct the trees, but using weighted versions of the original datasets. A set of trees are iteratively trained. When a tree is tested, the weights of each training records are updated in order to enable the subsequent tree to focus in the records that were misclassified by the previous tree. The predictions from all the trees are usually combined through a weighted majority vote. These weights are based on the accuracy of each tree.

In this case we have also constructed 100 trees.

3.4. *Overall academic performance*

As mentioned before, several studies have already explored the use of data mining tools to predict academic performance. However, the way academic performance is defined is not always the same.

GPA has been widely used as an indicator of academic performance because it is an objective measure with good internal reliability and temporal stability (e.g. [6,64]). However, it is not without limitations and several authors have explored other metrics to assess academic performance.

For example, Wati et al. [79]; Yue and Fu [81]; Miao and Haney [49] have explored time to degree completion. In fact, time to degree completion is a very important indicator for the institutions as it is linked to student and institutional success and accountability, education expenditure, and time investment [73].

Other studies have explored academic performance as the amount of study-points (credits) that students obtained in a certain point of their academic career [8,13,14]. Knowing that not all students enroll in the full program, and that frequent students fail specific courses and retake those courses the next academic year, Vanthournout et al. [78] propose not to use the absolute value of credits obtained but a ratio between the number of study credits a student obtained after a certain period of time and the total amount of credits the student was enrolled in during that period. In tune with this metric and with the purpose of also accommodating the primary indicator of academic performance, i.e. the grade point average [59] we propose a new metric of the overall academic performance of a student, i.e. the AP indicator.

The AP of a student at the end of the degree program (or in an advanced stage of the academic career) is defined by the ratio between the weighted mean of the grades obtained along the academic career (taking into account the number of ECTS credits the courses concluded are worth), and the total number of ECTS credits the student enrolled in to obtain approval. This performance indicator penalizes students with long academic careers and low average program grades.

$$AP = \frac{\sum_i G_i{*}C_i}{\sum_j C_j} \tag{1}$$

Eq. (1) presents the mathematical formulation of the proposed academic performance indicator. G_i represents the score obtained in each course i concluded with success, and C_i the corresponding number of ECTS credits. C_j refers to the number of ECTS credits of each course j a student enrolled in, to get approval. When a student fails a course, he or she has to enrol in that course as many times as needed to complete the course. Consequently, $\sum_j C_j$ is a proxy for the time required to conclude a degree or to achieve a certain stage of the academic career, as a student can enrol up to a limit number of ECTS credits in each academic year.

The reason for the introduction of the AP metric relies on the fact that academic performance is not only reflected by students' average grade but also by the effectiveness of the student during the academic path, and consequently the time taken to obtain approval in the curricular plan. This new approach captures, for example, in a situation in which two students have the same average grade, the one that did not have to repeat any course should have a higher performance score than the other that failed some courses and had to repeat them.

3.5. Evaluation criteria

In order to evaluate the performance of the techniques used to handle this problem, we use a validation procedure that aims to avoid overfitting. It is very important to address this issue, in order to quantify the ability of the model to generalize, i.e., its performance towards unseen data. We use a 10-fold cross validation, meaning that the data is divided into 10 blocks. The model is trained with 9 of the blocks, and the remaining one is used for testing purposes. The process is repeated 10 times, once for each of the different blocks. This validation procedure enables the maximization of the total number of observations used for testing.

In order to measure the performance of the proposed prediction model, we use the overall accuracy, the sensitivity (also called recall) and precision per class [80,39].

Consider the following multi-class confusion matrix (Fig. 3), referring to a three-class task {X, Y, Z}. We can define:

For example *a* refers to the number of observations where *X* was predicted and *X* was observed, while *b* refers to the number of observations that belong to *Y* and were predicted to belong to *X*.

The overall accuracy corresponds to the number of correct predictions divided by the total number of observations. Sensitivity refers to the fraction of observations of a class that were correctly predicted. Precision is defined as the fraction of correct predictions for a certain class.

4. Results and discussion

4.1. Performance level discretization

Based on the academic performance indicator introduced in Section 3.4, we use a discretization algorithm (equal-width binning) to establish five levels of academic performance. We decided to adopt the maximum number of clusters that is considered manageable by the institution.

The resulting performance levels are shown in Table 2. This table highlights the mean and standard deviation of the students' *AP* indicator included in each performance group.

For example, a student who is able to successfully conclude all the courses in which he enrolled for the first time, will have enrolled in a total of 300 ECTS by the end of the degree. Suppose that in half of the courses he received a final grade of 18 (out of 20), and in the other half of the courses he received a final grade of 16 (out of 20). In this scenario, he will have a weighted sum grade equal to 5100 (25 courses × 6 ECTS × 18 grade + 25 courses × 6 ECTS × 16 grade). This corresponds to an *AP* of 17 (=5100/300), and this student will be classified as a student A.

If another student also concludes all the courses, when enrolled for the first time, and gets a final grade of 14 (out of 20) in all the courses, he will be considered a student B (AP = 50 courses × 6 ECTS × 14 values/300 ECTS = 14).

		Observed		
		X	Y	Z
Predicted	X	a	b	c
	Y	d	e	f
	Z	g	h	i

Fig. 3. Confusion matrix.

Table 2
Performance levels for overall success.

Performance level	*AP* Mean (sd)	Percentage of students
A	16.8 (1.0)	10.33%
B	13.5 (1.1)	38.96%
C	9.8 (1.1)	28.14%
D	6.1 (1.1)	14.80%
E	2.5 (0.8)	7.77%

Consider another student who enrolled in courses corresponding to 384 ECTS during the student's academic path, meaning that the student had to repeat 14 courses, and received a final grade of 14 (out of 20) in all the courses. He is classified as a student C (AP = 50 courses × 6 ECTS × 14 values / 384 ECTS = 10.9).

We can observe in Table 2 that this classification results in a quite balanced share of students in each cluster.

4.2. Classification techniques performance

The performance of the proposed model was estimated when using each of the mentioned six classification techniques. Table 3 synthesizes the results obtained. The detailed results, including the confusion matrices obtained for each classification technique are introduced in Appendix A.3.

From the analysis of Table 3, we can conclude that the academic performance prediction in the beginning of the students' academic career is promising. Accuracy, precision and recall values are high when using each classification technique. This means that the institution in charge may adopt these techniques to predict the performance of new students. Random forests guarantees good prediction results, outperforming the other techniques in practically all explored performance metrics explored. This result corroborates the result in Zaklouta et al. [82], which refers that random forests achieve state-of-the-art performance in many multi-class classification applications. In opposition, and similarly to what happens in other settings (e.g. [9,27]), Naive Bayes is the worst performer, demonstrating the lowest values of all the considered metrics.

The other considered ensemble methods, i.e., bagging and adaptive boosting trees, also lead to accurate predictions. The results seem to reveal that, in this particular setting boosting decision trees generally outperform SVM, naive Bayes and decision trees. The superiority of the ensemble methods has been widely discussed in the literature [60].

In order to validate the results described, Table 4 presents the values obtained by performing a marginal homogeneity test for pairs of

Table 3
Classification techniques performance.

	Decision trees	SVM	Naive Bayes	Random forests	Bagging - decision trees	Adaptive boosting - decision trees
Accuracy	91.5%	93.9%	75.9%	96.1%	88.7%	95.7%
Recall A	91.0%	97.6%	80.9%	92.1%	75.2%	95.3%
Recall B	96.8%	95.2%	80.2%	98.5%	95.8%	96.5%
Recall C	91.5%	93.4%	71.5%	96.1%	86.6%	95.7%
Recall D	83.6%	89.3%	70.6%	93.4%	85.2%	95.1%
Recall E	81.0%	93.7%	73.5%	93.7%	85.9%	93.7%
Precision A	99.6%	88.9%	72.9%	99.2%	93.2%	95.3%
Precision B	91.1%	95.8%	81.0%	95.4%	84.2%	96.5%
Precision C	88.9%	93.8%	78.0%	94.1%	89.9%	95.9%
Precision D	90.7%	92.6%	68.1%	97.7%	93.7%	93.0%
Precision E	96.2%	95.2%	64.7%	100.0%	98.2%	96.8%

Table 4
Marginal homogeneity tests results (p-value).

	SVM	Naive Bayes	Random forests	Bagging - decision trees	Adaptive boosting - decision trees
Decision trees	8.11E-03	3.43E-13	**0.4630**	2.41E-02	**1.15E-01**
SVM		**1.78E-01**	4.76E-10	< 2.20E-16	7.64E-05
Naive Bayes			1.94E-3	4.60E-08	**9.28E-2**
Random forests				5.30E-12	3.70E-08
Bagging - decision trees					< 2.20E-16

techniques. It tests the equality of two multinomial response vectors. The null hypothesis is that the probability of being classified into one category is the same for the pair of techniques considered. The cells contain the p-value of the tests.

The results reveal that there is statistically significant difference between the models (significance level of 95%), with the exception of adaptive Boosting - decision trees and both decision trees and naive Bayes; random forests and decision trees; and SVM and naive Bayes.

4.3. Explanatory variables importance

We used random forests to derive the importance of each of the explanatory variables considered in the model (see Table 5), in order to get some insight into the importance of each feature in determining the academic success of a student. For this purpose, we followed the approach introduced by Menze et al. [48]. For each explanatory variable we considered all the nodes of the trees that used the variable under consideration and summed the increase in the accuracy promoted by the corresponding splits. Then we normalized the obtained sums, in order to bring all the values into the [0,1] range. Menze et al. [48] used Gini Index to evaluate the benefit promoted by each variable, while in this study we used accuracy for this purpose.

From the analysis of Table 5, it is interesting to note that the features that characterize the students in the enrolment process, particularly the enrolment average grade and the enrolment exams average grade, are the two most discriminating variables. This seems to suggest that the high school average grade is informative (8th position in the ranking), but that the combination with the enrolment exams average grade (resulting in the enrolment average grade) is more relevant. This is in line with the idea that there are differences between class-based and state-wide examination results [43]. High school final exams promote standardization in the grading system. Moreover, the relevance of enrolment average grade and the enrolment exams average grade attributes is aligned with a significant body of literature, that shows that both high school grade and standardized exams grades are generally strong predictors of student success in higher education institutions [17,38]. The second group of most informative variables is composed of those referring to the results obtained in the first academic year. However, it is worth noting that these features are considered less important than those previously highlighted, which may indicate that the results obtained in the first year itself are not sufficient to achieve high levels of accuracy in inferring students' performance. This fact may be linked to evidence that the first year in the university is a period of adjustment [52], influencing the results that students achieve. Focusing on the least informative attributes, the enrolment stage seems not to be very informative. This may be related to the fact that a very high majority of Engineering School students enrol in the first stage. The school type is also a weak predictor, in addition to socio-demographic aspects, such as gender and marital status. Regarding the middle positioned attributes, there are the socio-economic related attributes, i.e. parents' education and jobs. Despite not opposing the well-established positive relation between socio-economic status and academic achievement, the comparatively low position of the parents' education and job variables in the ranking, appears to suggest that, in this particular setting, the relationship is weaker than others that the model is able to capture.

4.4. Student segmentation and educational implications

We propose a students' segmentation framework, in order to easily and timely identify the target students for the different actions to be developed by the institutions for those specific segments. Therefore, we recommend the segmentation of students according to a cross-tabulation matrix that confronts the students' predicted performance levels with the different levels of academic success already observed in the first year of their studies (see Fig. 4). The indicator used to group students according to their academic achievement in the first year was the same used to group students by their final performance, i.e., the *AP* score.

Fig. 4 reveals that most of the students predicted to belong to one particular performance class by the end of the degree program, already give evidence of belonging to that performance class by the first year. In fact, the diagonal of the matrix refers to the majority of the students.

Table 5
Variables importance.

Explanatory variables	Type of variable	Normalized weight
Enrolment average grade	Enrolment process	1.00
Enrolment exams average grade	Enrolment process	0.79
Average grade in the first semester	First year assessment	0.76
Average grade in the second semester	First year performance	0.73
Number of ECTS completed in the second semester	First year performance	0.62
Number of ECTS completed in the first semester	First year performance	0.43
Average # of exams attended to conclude the courses in the first academic year	First year performance	0.40
High school average grade	High school background	0.39
Degree	Enrolment process	0.29
Mother's education level	Socio-economic status	0.26
Father's education level	Socio-economic status	0.26
Mother's job	Socio-economic status	0.23
Father's job	Socio-economic status	0.21
Academic year of enrolment	Enrolment process	0.17
Enrolment option	Enrolment process	0.17
Marital status	Socio-demographic	0.09
School type	High school background	0.06
Gender	Socio-demographic	0.05
Enrolment stage	Enrolment process	0.00

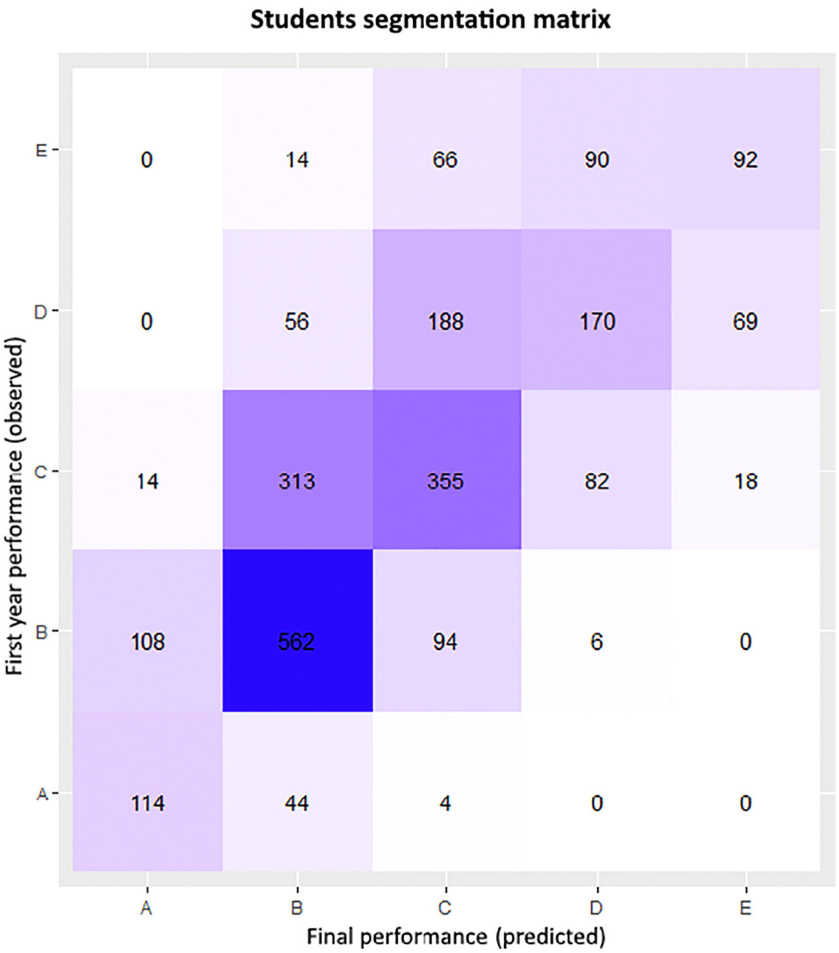

Fig. 4. Student segmentation matrix.

However, and as can be anticipated, the cells adjacent to the diagonal are very populated, meaning that there is a significant number of students whose performance levels achieved in the first year do not correspond to the final levels.

We believe that each segment of students, represented in each of these cells, should be targeted in a different way. Indeed, by using this approach, the institution may robustly sustain their differentiated actions, specifically targeted to mitigate or enhance performance effects. For example, if, in a certain moment, the institution is interested in involving those students who potentially are high achievers in a midterm tutoring system as tutors (e.g. [53]), the institution should focus on the 114 students who are predicted to belong to the top performers, and already demonstrate potential in their first year. If the institution did not consider their recent performance pattern, selecting, for example, the 14 students who in the first year were part of segment C, the peers would not recognize the students' value, and even the students in charge would not feel comfortable in that position. In opposition, if the institution would focus on the students who demonstrated high performance, regardless of the predicted performance, these students could potentially disappoint (e.g., the 4 students who ended up belonging to cluster C).

In another scenario, if, for example, the institution is interested in promoting the shift of those potentially lower performing students to higher levels of performance, by setting up a counselling program managed by the internal services (e.g. [71,41]), the institution should target those students whom the model predicts will have poor performance, and that already demonstrate signs of poor performance, i.e., the segment with 92 students. In fact, it may be difficult to persuade students who have not yet demonstrated difficulties to take part in the program. In opposition, to involve students who may naturally overcome their difficulties (for example, the 14 students who may end up belonging to cluster B) may constitute a waste of resources.

In another scenario, the institution may intend to address the 6 students that are grouped in class B in the first year, and whose predicting model classifies them as class D. This may reveal that if there is an appropriate follow-up, the student may overcome the potential tendency of lowering their performance.

5. Conclusions

This study proposes a model, supported by data mining classification techniques, that predicts students' overall academic performance based on the information available at the end of the first year of the students' academic path at the University.

This study confirms that prediction modelling is effective in the academic domain, and that decision makers can use such models to effectively plan the institution strategy and policy [51], as well as to optimize its limited resource allocation [19].

The results show that, in the context of the analysis, the information available at the end of the first year is sufficient to develop a solid model for a student's performance prediction. For example, the model supported by random forests was able to reach performance levels of about 96.1%, in terms of accuracy.

Moreover, the results reveal that, in this particular case study, random forests classification technique is the one presenting the best results, and naive Bayes is the one presenting the worst results. The results obtained also enable to conclude that the enrolment average grade and the enrolment exams average grade are the most relevant attributes for predicting the students' overall academic performance.

This study also proposes a students' segmentation framework, with the aim to distinguish students based on their observed achievements in the first year of the degree, and their propensity for academic success revealed by the prediction model. The proposed segmentation framework allows the identification of 25 segments of students and, thus, it may allow decision makers to specify different targets for the different actions to be developed.

From an educational standpoint, the authors believe that the proposed model is relevant, as it aids in inferring students' behaviour, and allows, for example, timely decision making for interventions that may either promote an increase in academic performance rates for the poor achievers or an increase in the quality of the students' academic experience for the high achievers.

On the one hand, using this tool, managers will be able to anticipate academic failure and act accordingly, in a proactive and timely manner, in order to change that tendency and mitigate its effects. For example, in order to stimulate those students whose propensity to not succeed is high and who have exhibited difficulties already in the first year of their academic path, decision makers could deploy measures that, for example, would involve psychological counselling services, the creation of tutoring systems and/or peer support groups, formal training in study skills, remedial lessons, etc. [35].

On the other hand, this tool can also be used to manage intellectual talent, and to enhance the impact of undergraduate education for particularly talented and motivated students, as it helps to identify potential academically-gifted students. Such students could either act as tutors for their peers (e.g. [53]), or become monitors/assistants in practical classes, motivating them for a future academic career, or set up honours programmes to attract talented students, providing them with the possibility to enhance their skills and pursue individual interests. Such programs may help schools compete for gifted students who, otherwise, may be attracted to other schools [31]. The accomplishments of these gifted students enhance universities' prestige and reputation, and leads to an atmosphere of academic vigour within the university community [65,32]. These programmes might also give students a slight advantage, in terms of grades earned and time to graduation [30].

6. Limitations and future work

This study suffers from some limitations. One of them is related to the sample bias. Despite considering a significant number of students, the conclusions drawn are only valid for the sample considered. In line with this, we only have data from a European institution. A good follow-up study would include other samples of data and would include other European as well as American faculties.

The model proposed was also tested with students attending only one learning area. Therefore, another possibility for future work is the extension of this project to other schools in different areas of learning, in which the students' performance patterns may be different. In fact, the adaptation and the application of the proposed method to other institutions may lead to different conclusions due to the heterogeneity observed among students of different scientific areas and of schools with different educational strategies.

Another limitation of the study is linked with the nature of the data considered. The prediction proposed model is based on data available in the enrollment period and is also based on the aggregated results achieved in the end of the first academic year. The possibility of predicting the academic performance of the students at the end of the first year using other data, such as questionnaires [64] and involvement data from - for example - moodle, constitutes an interesting topic for future research. Moreover, the use of results data for each course, instead of the aggregated results may be extremely relevant.

Regarding future work, the authors also believe that it is crucial to integrate the proposed method on the platforms already used by the institutions' educational decision makers, such as programme directors and committees, to support the development of early and appropriate educational measures, targeting the specific segmented groups of students.

Acknowledgements

We thank Mercedes Filho for the critical reading of the manuscript. We also thank the reviewers for the constructive reviews and comments on the manuscript.

Appendix A. Appendix

A.1. Related literature

Table A1
Studies addressing students' academic performance.

Study	Main objective	# Instances	Techniques	Dependent variable
Huang and Fang [35]	To predict student's scores on three dynamics mid-term exams	2907 exams/ students	Linear Regression, neural networks, support vector machines	Students' scores on the dynamics final comprehensive exam
Marbouti et al. [45]	To identify at-risk students in a course that used standards-based grading	2973 students	Logistic regression, neural networks, support vector machines, decision trees, naive Bayes, k-nearest neighbour	Fail or do not fail a course
Costa et al. [18]	To predict students likely to fail two courses (one performed on campus and another in a distance education format) at an early enough stage	262 + 161 students	Neural networks, support vector machines, decision trees, naive Bayes	Fail or do not fail a course
Gray et al. [26]	To identify college students at risk of failing in the first year of study	1193 students	Logistic regression, neural networks, support vector machines, decision trees, naive Bayes, k-nearest neighbour	First year poor academic achievers ($GPA < 2$) and strong academic achievers ($GPA > 2.5$)
Macfadyen and		118 students	Logistic regression	Fail or do not fail a course

(continued on next page)

Table A1 (*continued*)

Study	Main objective	# Instances	Techniques	Dependent variable
Dawson [44]	To identify which student online activities accurately predict academic achievement			
Guruler et al. [28]	To categorize students as either successful or unsuccessful and determine profiles of students whose GPA is equal to 2.0 (which is the minimum GPA required for graduation) or greater and of those students whose GPA is equal to 3.0 or greater (honor degree)	$>$ 2699 students	Decision trees	Final GPA is equal to 2.0 or greater + Final GPA is equal to 3.0 or greater
Laugerman et al. [40]	To determine what academic integration characteristics contribute to their success in engineering using post-hoc graduation data	472 students	Boosted regression	Earn or not an engineering degree
Hoffait and Schyns [34]	To identify freshmens' profiles likely to face major difficulties to complete their first academic year	6845 students	Logit regression, neural networks, random forests	Complete first year or not
Romero and Ventura [70]	To predict the marks that university students will obtain in the final exam of a course	438 + 135 + 438 students	Decision trees, neural networks, rule induction	Student fail in the exam of a course (value is < 5), pass (value is ≥ 5 and < 7), has a good mark (value is ≥ 7 and < 9) or has an excellent mark (value is ≥ 9)
Palmer [54]	To predict academic performance of engineering students enrolled in a second-year class	132 students	Logistic regression	Fail or do not fail a course
Romero et al. [66]	To predict students' final marks based on their participation in forums	114 students	Logit regression, neural networks, random forests, naive Bayes, BayesNet, Sequential Minimal Optimization	Fail or do not fail a course
Mishra et al. [50]	To predict the third semester performance of MCA students	250 students	Decision trees, random trees	Third semester performance as BAVG ($<$ 60%), AVG (60% to less than 70%), ABVG (70% to less than 79%) and EXCL ($\geq$ 80%)
Natek and Zwilling [51]	To predict the success rate of students enrolled in their courses	106 students	Decision trees	High (values between 8 and 10), medium (values between 6 and 7), low final grade (values lower than 6) in informatics courses
Arsad et al. [5]	To predict the academic performance of Electrical Degree students	391 + 505 students	Neural Networks	Cumulative Grade Point Average (CGPA) in semester 8
Strecht et al. [74]	To predict approval/failure in a course and to predict its grade	5779 courses/ students	Linear regression, neural networks, support vector machines, decision trees, naive Bayes, k-nearest neighbour, adaBoost	Fail or do not fail a course; Final Grade
Vandamme et al. [77]	To predict the first year performance of students	227 students	Discriminant analysis, neural networks, random forests and decision trees	High risk (average mark of less than 45% in a session), Low risk (average of more than 70% a session), Medium risk (average grade between 45% and 70%)
Aluko et al. [2]	To predict academic success of architecture students based on information provided in prior academic performance	101 students	Discriminant analysis, k-nearest neighbour	Pass (CGPA at graduation between 5 and 2.4) and Fail (CGPA at graduation between 2.39 and 0)
Current study	To predict performance levels in the end of the degree (or at an advanced stage of the academic career), in the end of the first academic year	2459 students	Random forests, decision trees, support vector machines, naive Bayes, bagged trees and boosted trees	A, B, C, D, E levels determined based on the ratio between the weighted mean of the grades obtained along the academic career and the total number of ECTS credits the student enrolled in

A.2. Dataset variables

Table A2
Dataset variables.

Attribute	Type	Description	Values(frequency)/Mean(std. deviation)
1	Categorical	Weighted mean of the grade	A:10.3% B:39.0% C:28.1% D:14.8% E:7.8%
2	Categorical	Gender	Male:77.5% Female:22.5%
3	Categorical	Marital status	Single:98.0% Married:2.0%
4	Categorical	Father's education level	Undergraduate level:20.2% Unknown:19.4% 1st cycle of elementary school:13.9% Secondary education:13.4% 3rd Cycle of Elementary School:12.7% 2nd Cycle of Elementary School:6.8% Bachelors:4.1% Postgraduate level (masters):3.3% Postgraduate level (PhD):2.4% Postsecondary non-higher education:1.8% Able to read and write but has not completed 1st cycle of elementary school:1.1% Technological specialization (higher education):0.8% Not able to read and write:0.1%
5	Categorical	Father's job	Unknown:27.8% Other activity:17.6% Qualified civil servants, directors and qualified company employees:13.3% Intermediate-level technicians and professionals:9.1% Specialists in intellectual and scientific professions:7.6% Services and sales staff:6.9% Workers, craftsmen and similar:6.5% Administrative staff and similar workers:4.5% Unskilled workers:2.6% Machine operators and assembly workers:1.6% Farmers and workers qualified in agriculture and fishing:1.4% Members of the armed forces:1.1%
6	Categorical	Mother's education level	Undergraduate level:26.6% Unknown:18.8% 3rd cycle of elementary school:12.2% Secondary education:11.2% 1st Cycle of Elementary School:10.0% 2nd Cycle of Elementary School:7.6% Postgraduate level (masters):3.6% Able to read and write but has not completed 1st cycle of elementary school:3.3% Bachelors:3.2% Postgraduate level (PhD):1.5% Postsecondary non-higher education:1.5% Technological specialization (higher education):0.4% Not able to read and write:0.1%
7	Categorical	Mother's job	Unknown:34.9% Other activity:19.2% Qualified civil servants, directors and qualified company employees:10.2% Specialists in intellectual and scientific professions:9.6% Administrative staff and similar workers:8.3% Intermediate-level technicians and professionals:5.9% Services and sales staff:4.8%

(*continued on next page*)

Table A2 (*continued*)

Attribute	Type	Description	Values(frequency)/Mean(std. deviation)
			Unskilled workers:3.2% Workers, craftsmen and similar:3.0% Farmers and workers qualified in agriculture and fishing:0.6% Machine operators and assembly workers:0.3%
8	Categorical	School type	Public:83.7% Private:16.3%
9	Categorical	Enrolment option	1: 81.5% 2: 12.3% 3: 3.7% 4: 1.6% 5: 0.6% 6: 0.3%
10	Categorical	Enrolment stage	1: 87.3% 2: 12.7%
11	Categorical	Academic year of enrolment	2003:23.8% 2004:21.7% 2005:16.0% 2006:18.2% 2007:20.3%
12	Numerical	High school average grade	16.0 (1.5)
13	Numerical	Enrolment exams average grade	148.4 (22.3)
14	Numerical	Enrolment average grade	154.3 (16.3)
15	Categorical	Degree	Civil engineering (LEC):17.5% Electrical and Computers Engineering (LEEC):14.5% Informatics and Computing Engineering (LIEC):9.6% Master in Civil Engineering (MIEC):9.6% Mechanical Engineering (LEM):9.2% Master in Electrical and Computers Engineering (MIEEC):7.7% Master in Informatics and Computing Engineering (MIEIC):6.4% Master in Mechanical Engineering (MIEM):5.9% Industrial Engineering and Management (LGEI):4.1% Chemical Engineering (LEQ):3.4% Master in Environmental Engineering (MIEA):3.1% Master in Industrial Engineering and Management (MIEIG):1.9% Master in Chemical Engineering (MIEQ):1.9% Engineering and Environment Management (LEGA):1.5% Master in Bioengineering (MIB):1.5% Metallurgical and Materials Engineering (LEMM):1.5% Master in Metallurgical and Materials Engineering (MIEMM):0.7%
15	Numerical	Average number of exams attended to conclude the courses of the first year	6.8 (1.5)
16	Numerical	Number of ECTS completed in the first semester	24.6 (6.9)
17	Numerical	Average grade in the first semester	13.14 (1.6)
18	Numerical	Number of ECTS completed in the second semester	22.5 (8.3)
19	Numerical	Average grade in the second semester	12.6 (1.7)

A.3. Confusion matrices

Table A3
Confusion matrix - Decision trees.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	233	1	0	0	0	1.00
	B	22	929	46	21	2	0.91

(*continued on next page*)

Table A3 (*continued*)

		Observed					
		A	B	C	D	E	Prec.
	C	1	21	635	34	23	0.89
	D	0	9	11	301	11	0.91
	E	0	0	2	4	153	0.96
	Recall	0.91	0.97	0.91	0.84	0.81	

Table A4
Confusion matrix - SVM.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	248	24	3	4	0	0.89
	B	6	912	26	8	0	0.96
	C	0	18	646	20	5	0.94
	D	0	4	15	325	7	0.93
	E	0	0	2	7	179	0.95
	Recall	0.98	0.95	0.93	0.89	0.94	

Table A5
Confusion matrix - NB.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	207	68	9	0	0	0.73
	B	44	770	102	31	4	0.81
	C	5	73	496	44	18	0.78
	D	0	38	53	254	28	0.68
	E	0	11	34	31	139	0.65
	Recall	0.81	0.80	0.71	0.71	0.74	

Table A6
Confusion matrix - RF.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	234	2	0	0	0	0.99
	B	20	944	19	2	4	0.95
	C	0	12	665	22	8	0.94
	D	0	0	8	340	0	0.98
	E	0	0	0	0	179	1.00
	Recall	0.92	0.99	0.96	0.93	0.94	

Table A7
Confusion matrix - Bagging DT.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	191	12	2	0	0	0.93

(*continued on next page*)

Table A7 (*continued*)

		Observed					
		A	B	C	D	E	Prec.
	B	60	918	85	23	4	0.84
	C	3	21	599	30	13	0.90
	D	0	7	4	310	10	0.94
	E	0	0	2	1	164	0.98
	Recall	0.75	0.96	0.87	0.85	0.86	

Table A8
Confusion matrix - Boosting DT.

		Observed					
		A	B	C	D	E	Prec.
Predicted	A	242	12	0	0	0	0.95
	B	10	924	14	6	4	0.96
	C	0	18	662	10	0	0.96
	D	2	2	14	346	8	0.93
	E	0	2	2	2	179	0.97
	Recall	0.95	0.96	0.96	0.95	0.94	

References

[1] S.B. Aher, L.M.R.J. Lobo, Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data, Knowledge-Based Systems 51 (2013) 1–14.

[2] R.O. Aluko, O.A. Adenuga, P.O. Kukoyi, A.A. Soyingbe, J.O. Oyedeji, Predicting the academic success of architecture students by pre-enrolment requirement: using machine-learning techniques, Construction Economics and Building 16 (4) (2016) 86–98.

[3] American Council on Education, First in the World by 2020: What Will It Take? http://www.acenet.edu/the-presidency/columns-and-features/Pages/First-in-the-World-by-2020-What-Will-It-Take.aspx, (2011) [Online; accessed 14-July-2017].

[4] F. Araque, C. Roldán, A. Salguero, Factors influencing university drop out rates, Computers & Education 53 (3) (2009) 563–574.

[5] P.M. Arsad, N. Buniyamin, J.l.A. Manan, A neural network students' performance prediction model (NNSPPM), 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), 2013, pp. 1–5.

[6] D.R. Bacon, B. Bean, GPA in research studies: an invaluable but neglected opportunity, Journal of Marketing Education 28 (1) (2006) 35–42.

[7] R. Baeza-Yates, Z. Liaghat, Quality-efficiency trade-offs in machine learning for text processing, 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 897–904.

[8] M.N.V.D. Berg, W.H.A. Hofman, Student success in university education: a multi-measurement study of the impact of student and faculty factors on study progress, Higher Education 50 (3) (2005) 413–446.

[9] M. Bogaert, M. Ballings, D. Van den Poel, The added value of Facebook friends data in event attendance prediction, Decision Support Systems 82 (2016) 26–34.

[10] G. Bordea, A.M. Shahiri, W. Husain, N.A. Rashid, A review on predicting students performance using data mining techniques, Procedia Computer Science 72 (2015) 414–422.

[11] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.

[12] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.

[13] V.V. Busato, F.J. Prins, J.J. Elshout, C. Hamaker, Learning styles: a cross-sectional and longitudinal study in higher education, British Journal of Educational Psychology 68 (3) (1998) 427–441.

[14] V.V. Busato, F.J. Prins, J.J. Elshout, C. Hamaker, Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education, Personality and Individual Differences 29 (6) (2000) 1057–1068.

[15] H. Bydzovska, Course enrollment recommender system, in: T. Barnes, M.F. Min Chi (Eds.), Proceedings of the 9th International Conference on Educational Data Mining, International Educational Data Mining Society, Raleigh, NC, USA, 2016, pp. 312–317.

[16] T.M. Christian, M. Ayub, Exploration of classification using NBTree for predicting students' performance, 2014 International Conference on Data and Software Engineering (ICODSE), 2014, pp. 1–6, , https://doi.org/10.1109/ICODSE.2014.7062654.

[17] E. Cohn, S. Cohn, D.C. Balch, J. Bradley, Determinants of undergraduate GPAs: SAT scores, high-school GPA and high-school rank, Economics of Education Review 23 (6) (2004) 577–586.

[18] E.B. Costa, B. Fonseca, M.A. Santana, F.F. de Araújo, J. Rego, Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, Computers in Human Behavior 73 (2017) 247–256.

[19] D. Delen, A comparative analysis of machine learning techniques for student retention management, Decision Support Systems 49 (4) (2010) 498–506.

[20] T.G. Dietterich, Ensemble methods in machine learning, Multiple Classifier Systems, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2000, pp. 1–15.

[21] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: A. Prieditis, S. Russell (Eds.), Machine Learning Proceedings 1995, Morgan Kaufmann, San Francisco (CA), 1995, pp. 194–202.

[22] A.O. Elfaki, K.M. Alhawiti, Y.M. AlMurtadha, O.A. Abdalla, A.A. Elshiekh, Supporting students' learning-pathway choices by providing rule-based recommendation system, International Journal of Education and Information Technologies 9 (2015) 81–94.

[23] European Commission, Europe 2020 targets. http://ec.europa.eu/eurostat/web/europe-2020-indicators/europe-2020-strategy/targets, (2010) [Online; accessed 9-October-2017].

[24] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, In Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann, 1996, pp. 148–156.

[25] D. Gibson, Big data in higher education: research methods and analytics supporting the learning journey, Technology, Knowledge and Learning 22 (3) (2017) 237–241.

[26] G. Gray, C. McGuinness, P. Owende, An application of classification models to predict learner progression in tertiary education, Advance Computing Conference (IACC), 2014 IEEE International, 2014, pp. 549–554.

[27] S.S. Groth, J. Muntermann, An intraday market risk management approach based on textual analysis, Decision Support Systems 50 (4) (2011) 680–691.

[28] H. Guruler, A. Istanbullu, M. Karahasan, A new student performance analysing system using knowledge discovery in higher educational databases, Computers & Education 55 (1) (2010) 247–254.

[29] J.J.B. Harlow, D.M. Harrison, A. Meyertholen, Correlating student interest and high school preparation with learning and performance in an introductory university physics course, Physical Review Special Topics-Physics Education Research 10 (1) (2014) 010112.

[30] K.A. Harper, L. Abrams, J.P. Ruffley, A Longitudinal Study of the Impact of a First-Year Honors Engineering Program, 2014, pp. 1–8.

[31] J.C. Hartshorn, V.A. Berbiglia, M. Heye, An honors program: directing our future leaders, The Journal of Nursing Education 36 (4) (1997) 187–189.

[32] T.P. Hébert, M.T. McBee, The impact of an undergraduate honors program on gifted university students, Gifted Child Quarterly 51 (2) (2007) 136–151.

[33] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf, Support vector machines, IEEE Intelligent Systems and their Applications 13 (4) (1998) 18–28.

[34] A.-S. Hoffait, M. Schyns, Early detection of university students with potential difficulties, Decision Support Systems 101 (2017) 1–11.

[35] S. Huang, N. Fang, Predicting student academic performance in an engineering dynamics course: a comparison of four types of predictive mathematical models,

Computers & Education 61 (2013) 133–145.
[36] N. Jeeva, Elakia, Gayathri, Aarthi, Application of data mining in educational database for predicting behavioural patterns of the students, International Journal of Computer Science and Information Technologies 5 (3) (2014) 4469–4472.
[37] M.Y. Kiang, D.M. Fisher, J.-C.V. Chen, S.A. Fisher, R.T. Chi, The application of SOM as a decision support tool to identify AACSB peer schools, Decision Support Systems 47 (1) (2009) 51–59.
[38] N.R. Kuncel, M. Crede, L.L. Thomas, D.M. Klieger, S.N. Seiler, S.E. Woo, A meta-analysis of the validity of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student performance, American Journal of Pharmaceutical Education 69 (3) (2005) 51.
[39] V. Labatut, H. Cherifi, Accuracy measures for the comparison of classifiers, in: A.-D. Ali (Ed.), The 5th International Conference on Information Technology, Al-Zaytoonah University of Jordan, amman, Jordan, 2011, p. 1,5.
[40] M. Laugerman, D. Rover, M. Shelley, S. Mickelson, Determining graduation rates in engineering for community college transfer students using data mining, International Journal of Engineering Education (2015) 1448–1457.
[41] D. Lee, E.A. Olson, B. Locke, S.T. Michelson, E. Odes, The effects of college counseling services on academic performance and retention, Journal of College Student Development 50 (3) (2009) 305–319.
[42] D.D. Lewis, Naive (Bayes) at forty: the independence assumption in information retrieval, in: C. Nédellec, C. Rouveirol (Eds.), Machine Learning: ECML-98, number 1398 in Lecture Notes in Computer Science, Springer Berlin Heidelberg, 1998, pp. 4–15.
[43] K. Maag Merki, Effects of the implementation of state-wide exit exams on students' self-regulated learning, Studies in Educational Evaluation 37 (4) (2011) 196–205.
[44] L.P. Macfadyen, S. Dawson, Mining LMS data to develop an early warning systemİ for educators: a proof of concept, Computers & Education 54 (2) (2010) 588–599.
[45] F. Marbouti, H.A. Diefes-Dux, K. Madhavan, Models for early prediction of at-risk students in a course using standards-based grading, Computers & Education 103 (2016) 1–15.
[46] C. Márquez-Vera, A. Cano, C. Romero, A.Y.M. Noaman, H. Mousa Fardoun, S. Ventura, Early dropout prediction using data mining: a case study with high school students, Expert Systems 33 (1) (2016) 107–124.
[47] M. Mayilvaganan, D. Kalpanadevi, Comparison of classification techniques for predicting the performance of students academic environment, 2014 International Conference on Communication and Network Technologies (ICCNT), 2014, pp. 113–118, , https://doi.org/10.1109/CNT.2014.7062736.
[48] B.H. Menze, B.M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, F.A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics 10 (2009) 213.
[49] J. Miao, W. Haney, High school graduation rates:alternative methods and implications, Education Policy Analysis Archives 12 (0) (2004) 55.
[50] T. Mishra, D. Kumar, S. Gupta, Mining students' data for prediction performance, 2014 Fourth International Conference on Advanced Computing Communication Technologies, 2014, pp. 255–262.
[51] S. Natek, M. Zwilling, Student data mining solution-knowledge management system related to higher education institutions, Expert Systems with Applications 41 (14) (2014) 6400–6407.
[52] S.M. Nightingale, S. Roberts, V. Tariq, Y. Appleby, L. Barnes, R.A. Harris, L. Dacre-Pool, P. Qualter, Trajectories of university adjustment in the United Kingdom: emotion management and emotional self-efficacy protect against initial poor adjustment, Learning and Individual Differences 27 (2013) 174–181.
[53] O. Nomura, H. Onishi, H. Kato, Medical students can teach communication skills - a mixed methods study of cross-year peer tutoring, BMC Medical Education 17 (2017) 103.
[54] S. Palmer, Modelling engineering student academic performance using academic analytics, International journal of engineering education 29 (1) (2013) 132–138.
[55] Z.K. Papamitsiou, V. Terzis, A.A. Economides, Temporal learning analytics for computer based testing, Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, LAK '14, ACM, New York, NY, USA, 2014, pp. 31–35.
[56] S. Parack, Z. Zahid, F. Merchant, Application of data mining in educational databases for predicting academic trends and patterns, 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), 2012, pp. 1–4.
[57] A. Pena-Ayala (Ed.), Educational Data Mining, volume 524 of Studies in Computational Intelligence, Springer International Publishing, Cham, 2014.
[58] A. Pena-Ayala, Educational data mining: a survey and a data mining-based analysis of recent works, Expert Systems with Applications 41 (4) (2014) 1432–1462.
[59] G.R. Pike, J.L. Saupe, Does high school matter? An analysis of three methods of predicting first-year grades, Research in Higher Education 43 (2) (2002) 187–207.
[60] S. Piri, D. Delen, T. Liu, H.M. Zolbanin, A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble, Decision Support Systems 101 (Supplement C) (2017) 12–27.
[61] G. Putnik, E. Costa, C. Alves, H. Castro, L. Varela, V. Shah, Analysing the correlation between social network analysis measures and performance of students in social network-based engineering education, International Journal of Technology and Design Education 26 (3) (2016) 413–437.
[62] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[63] Rapidminer, Rapidminer, https://rapidminer.com/, (2018) [Online; accessed 14-Mar-2017].
[64] M. Richardson, C. Abraham, R. Bond, Psychological correlates of university students' academic performance: a systematic review and meta-analysis, Psychological Bulletin 138 (2) (2012) 353–387.
[65] A.N. Rinn, J.A. Plucker, We recruit them, but then what? The educational and psychological experiences of academically talented undergraduates, Gifted Child Quarterly 48 (1) (2004) 54–67.
[66] C. Romero, P.G. Espejo, A. Zafra, J.R. Romero, S. Ventura, Web usage mining for predicting final marks of students that use Moodle courses, Computer Applications in Engineering Education 21 (1) (2013) 135–146.
[67] C. Romero, M.-I. López, J.-M. Luna, S. Ventura, Predicting students' final performance from participation in on-line discussion forums, Computers & Education 68 (2013) 458–472.
[68] C. Romero, S. Ventura, Educational data mining: a survey from 1995 to 2005, Expert Syst. Appl. 33 (1) (2007) 135–146.
[69] C. Romero, S. Ventura, Educational data mining: a review of the state of the art, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 40 (6) (2010) 601–618.
[70] C. Romero, S. Ventura, Data mining in education, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 3 (1) (2013) 12–27.
[71] J. Russell, G. Thomson, D. Rosenthal, International student use of university health and counselling services, Higher Education 56 (1) (2008) 59–75.
[72] I.A. Saad, Predictive validity of high school performance with respect to academic achievement at the university level, International Journal of Psychology 43 (3-4) (2008) 772-772.
[73] D. Shapiro, A. Dundar, P.K. Wakhungu, X. Yuan, A. Nathan, Y. Hwang, Time to Degree: A National View of the Time Enrolled and Elapsed for Associate and Bachelors Degree Earners, Technical report, Herndon, VA: National Student Clearinghouse Research Center, 2016.
[74] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, R. Abreu, A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance, International Educational Data Mining Society, Madrid, 2015.
[75] D. Thammasiri, D. Delen, P. Meesad, N. Kasap, A critical assessment of imbalanced class distribution problem: the case of predicting freshmen student attrition, Expert Systems with Applications 41 (2) (2014) 321–330.
[76] V. Tinto, Limits of theory and practice in student attrition, The Journal of Higher Education 53 (6) (1982) 687–700.
[77] J.-P. Vandamme, N. Meskens, J.-F. Superby, Predicting academic performance by data mining methods, Education Economics 15 (4) (2007) 405–419.
[78] G. Vanthournout, D. Gijbels, L. Coertjens, V. Donche, P. Van Petegem, Students' persistence and academic success in a first-year professional bachelor program: the influence of students' learning strategies and academic motivation, Education Research International 2012 (2012) 1–10.
[79] M. Wati, Haeruddin, W. Indrawan, Predicting degree-completion time with data mining, 2017 3rd International Conference on Science in Information Technology (ICSITech), 2017, pp. 732–736, , https://doi.org/10.1109/ICSITech.2017.8257209.
[80] I.H. Witten, E. Frank, M.A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann, USA, 2011.
[81] H. Yue, X. Fu, Rethinking graduation and time to degree: a fresh perspective, Research in Higher Education 58 (2) (2017) 184–213.
[82] F. Zaklouta, B. Stanciulescu, O. Hamdoun, Traffic sign classification using K-d trees and Random Forests, The 2011 International Joint Conference on Neural Networks (IJCNN), 2011, pp. 2151–2155.

V. L. Miguéis is an Assistant Professor in the department of Industrial Engineering and Management at the School of Engineering of the University of Porto, Portugal. She received her PhD in Industrial Engineering and Management from the School of Engineering of the University of Porto. Her research interests include educational mining, customer relationship management, data mining, customer intelligence and forecasting. Her research specifically focuses on the use of data mining techniques to support the decision process. She has published papers in several international journals indexed in the Web of Knowledge. She has taught courses in operations research, data mining, statistics and operations management. She is the external relations manager of the Industrial Engineering and Management Master of the School of Engineering of the University of Porto.

Ana Freitas has a degree in Education and a Master in Educational Sciences, both from the University of Minho, Portugal. She's also a PhD candidate in Educational Sciences at the University of Porto, Portugal. She's been working, for 15 years, in Higher Education Institutions, in the area of educational consultancy & support, academic management and research in education. She is currently working as a Senior Officer at the Laboratory of Teaching and Learning, in the School of Engineering, University of Porto. Her specific interest areas and field of activities are in professors' continuing professional development, doctoral education/innovative doctoral training and academic success. She collaborated in 6 financed projects. She's accredited as a trainer of trainers and as a trainer of teachers/professors.

Paulo J. V. Garcia holds a PhD in Physics from Université de Lyon (France). He is an Associate Professor at the Department of Engineering Physics at the School of Engineering of the University of Porto, Portugal and is also the Coordinator of the Teaching Learning Laboratory of the same school. He conducts research mostly in the fields of astrophysics, optical instrumentation, signal processing and physics education.

André Silva received his Master's degree in Informatics and Computing Engineering from the School of Engineering of the University of Porto, Portugal. He became interested in educational mining and business analytics after having been developing his masters' project.