MOX-Report No. 41/2020

Not the magic algorithm: modelling and early-predicting students dropout through machine learning and multilevel approach

Cannistrà,M.; Masci, C.; Ieva, F.; Agasisti, T.; Paganoni, A.M.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Not the magic algorithm: modelling and early-predicting students dropout through machine learning and multilevel approach

M. Cannistrà[#], C. Masci[♮], F. Ieva[♮], T. Agasisti[#], and A. M. Paganoni[♮]

June 4, 2020

[#] DIG - Department of Management Engineering,
Politecnico di Milano, via Lambruschini 4/b, Milano, Italy
tommaso.agasisti@polimi.it
marta.cannistra@polimi.it
[♮] MOX - Modelling and Scientific Computing, Department of Mathematics,
Politecnico di Milano, via Bonardi 9, Milano, Italy
chiara.masci@polimi.it
anna.paganoni@polimi.it
ieva.francesca@polimi.it

Abstract

According to OECD, almost 30 per cent of students leave tertiary education programs without obtaining a degree. This number measures a dead loss of human capital and a waste of public and private resources. This paper contributes to the existing knowledge about students dropout by combining a theoretical-based model with a data-driven approach to detect students who are more likely to leave university in the first year. We propose the use of multilevel statistical models and machine learning methods, applied to administrative data from a leading Italian university. The findings are encouraging, as the methodology is able to predict at-risk students very precisely. We provide evidence of the essential role of data relative to early performance (i.e. grades obtained in the first semester). Moreover, the selection of major strongly influences the probability of dropping out.

Keywords: Learning Analytics, Early Warning Systems, Student dropout, Machine Learning, multilevel models, HE students.

> "Universities should use data regularly and systematically to identify high-risk students, target them with interventions, and evaluate those interventions' effectiveness"
>
> von Hippel and Hofflinger, 2020

1 Introduction

The Italian Higher Education (HE) system is plagued by a high level of dropout, with many students abandoning their Bachelor courses during the first or second year. According to the Italian National Agency for the Evaluation of Universities and Research Institutes (ANVUR), the dropout rate for the cohort of students from whom complete data are available is around 28.2 per cent, with almost two-thirds of them (20 per cent) dropping out in the first two years ("ANVUR: Rapporto biennale sullo stato del Sistema Universitario e della ricerca", 2018). This data is particularly worrying because of the low proportion of people holding a tertiary education degree in Italy. OECD (2019) indicates that the percentage of 25-34 years old adults with higher education was 28 per cent, with the same share being 19 per cent for the adults 25-64 years old (reference year: 2018) - both indicators are well below the OECD average. Understanding the dropout phenomenon is so important, in Italy, that a number of academic studies explored it under many different viewpoints. For example, Belloc, Maruotti, and Petrella (2010) aims at individuating personal features of students who are more likely to dropout (instead of universities' institutional factors), administering a questionnaire directly to the students. In the related study in Belloc, Maruotti, and Petrella (2011), the authors utilise administrative data from one university for the same purpose, employing novel statistical techniques in the analysis. The authors in Aina (2013) use (Italian) data from the European Community Household Panel and detects the strong role of parental background in affecting persistence - with students from disadvantaged families more likely to dropout, all other factors held constant. Such difference related with socioeconomic background has been confirmed more recently by Ghignoni (2017) as well as by Contini, Cugnata, and Scagni (2018). The work in Di Pietro and Cutillo (2008) suggests that degree flexibility can help reducing likelihood to dropout - and employ data about a national reform undertaken in Italy 2001 for testing this intuition. A high incidence of dropout rates in the functioning of the HE system generates equity and efficiency issues. On the equity side, various students demonstrate how there is a correlation between socioeconomic background and dropout, and the academic literature confirms that disadvantaged students are more at-risk of dropping out. Unfortunately, reforms and interventions for expanding the access to HE were not successful in reducing the socioeconomic gradient of the dropout (Bratti, Checchi, & De Blasio, 2008; Brunori, Peragine, & Serlenga, 2012; Oppedisano, 2011). When considering efficiency, dropout represents a net waste of resources. Indeed, educating students is a costly activity, which generates returns in the long run due to the credentials acquired and the human capital accumulated. When students do not conclude their courses with a degree, these benefits are not realised and only the costs accrue to the educational activities. A recent trend in the interventions for improving retention and reducing dropout rates is the use of Learning Analytics tools (De Freitas et al., 2015). Specifically, the use of advanced techniques, rooted in both the statistical and Machine Learning domains, is applied to predict the students who are more at-risk of dropping out. If algorithms demonstrate to be effective in predicting students' performance, the early identification of students at-risk can be helpful for designing targeted interventions for improving their chances of retention (Burgos et al., 2018). While a growing number of studies starts considering the specific use of predictions for remedial education, the debate about the best models to be employed for predictions is far from

being concluded, and the empirical solutions proposed are not widely accepted. In this paper, we use administrative data from Politecnico di Milano (PoliMi), Italy, to test some novel models to formulate predictions of at-risk students. The database gathers various cohorts of first-year Bachelor students (in Engineering) and covers 9 years (from 2010 to 2019); overall, it includes more than 110,000 students, with associated 10,000,000 entries, each of which is a specific event related with the student journey (her initial administrative record, exams, etc.). The research stems from an institutional initiative launched by PoliMi under the label "Data Analytics for Institutional Support", which broad aim is to leverage the available (administrative) datasets of the university to analyze many aspects of the academic life, and support better decision-making. The priority assigned to the Research Group is to detect the main causes of students' dropout, which is substantial in PoliMi (about 30 per cent). At the same time, the initiative's objective is to create remedial initiatives, for helping students at-risk to avoid giving-up. The primary step consisted in the development of an algorithm that should be able to identify students at-risk early in their academic career, i.e. at the end of the 1^{st} semester of the 1^{st} year. The Research Group performed various analyses for setting the best performing algorithms as possible, embarking a profound methodological work along with deep empirical testing. This fundamental step will pave the way to subsequent interventions for reducing dropout and for testing their effectiveness, which represents the most recent frontier of the academic literature on this specific use of Learning Analytics (Larrabee Sønderlund, Hughes, & Smith, 2019). In this work, we report the findings from the application of newly developed algorithms to the problem of early detecting students who are potentially at-risk of dropout. We use various cohorts of 1^{st} year Bachelor students as the applicative case, including Engineering students only (thus, excluding Architecture and Design). This paper answers three research questions:

a. How important is that the predictive algorithm takes into account the grouped nature of data, i.e. considering that students are enrolled to different degree programs - so that their probability of dropping out is conditional to the degree program they chose?

b. Which characteristics are more frequently associated to the risk of dropping out? And specifically, how important are the different groups of variables that are available in improving prediction? - please note that we have data about individual demographics, prior achievement and current academic results.

c. How do alternative algorithms' types (Machine Learning vs generalised linear models) perform in predicting actual dropout?

This paper innovates the current state-of-the-art of the field in two main directions. First, we develop a comprehensive theoretical model for studying dropout in a data analysis perspective, complementing the application of techniques to the existing data with a conceptual approach for exploring the determinants of dropout. The current approaches based on Learning Analytics are indeed very much data-driven, while paying less attention to the theoretical foundations of the models developed for the empirical analyses. We build a bridge between the literature about university dropout/success (Aljohani, 2016) and the one about the use of Learning Analytics techniques in the field (Daniel, 2015; Leitner, Khalil, & Ebner, 2017). Second, we compare different algorithms, built following alternative hypotheses and specifications, to test the validity and robustness of a number of statistical and Machine Learning methods. Given that the practical use and application of predictive models will produce real effects on the academic life of students (for example, by activating targeted interventions), assessing the reliability of the algorithms and their stability across specifications becomes a crucial feature of any exercise of this kind. The remainder of the paper is organised as follows. In the section 2, we

develop the theoretical framework for deriving the empirical models in the Learning Analytics perspective. Section 3 describes the available data, together with a background about the main characteristics of PoliMi - which are necessary to put the case in perspective and it illustrates the methodologies employed for the empirical analysis. Section 4 reports the main results, ordered according to the three research questions. Lastly, Section 5 discusses the main implications and general suggestions towards implementing future interventions for helping at-risk students.

2 Theoretical framework

2.1 The individual educational timeline - an overview

The investigation of dropout phenomenon within Higher Education Institutions (HEIs) has been a concern for educators, university managers and policy makers. The academic literature distinguishes between two approaches investigating the features of this phenomenon: theory-driven and data-driven. The first stream deepens the reasons and the psychological constructs behind withdrawing decisions, identifying theoretical fundamentals and drawing a conceptual model to guide the inquiry. Different authors (Cabrera, Stampen, & Hansen, 1990; John, Paulsen, & Starkey, 1996; Pascarella & Terenzini, 1980; Spady, 1970; Tinto, 1975) propose their models to show the processes of interactions between students, their characteristics and the institutions that lead to dropout (Tinto, 1975). In particular, the model considers the interaction between the student and the university environment in which individual attributes are exposed to influences, expectations, and demands from a variety of sources (such as courses, faculty members, administrators, and peers). The interaction between these two aspects allows the student to have success or failure in both the academic and social system (Spady, 1970). Hence, these studies focus on the necessity to contextualise the student's educational career in a specific community and situation.

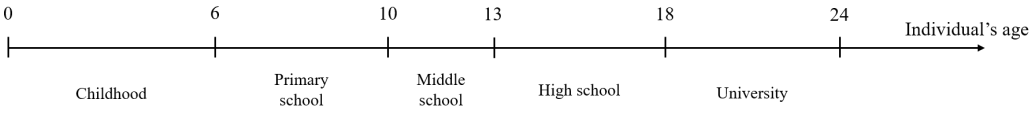
An alternative approach deals with data-driven studies, in which students' characteristics are analysed longitudinally to find the best statistical models predicting dropout or graduation (Barbu et al., 2019; Korhonen & Rautopuro, 2019; Li, Rusk, & Song, 2013; Seidel & Kutieleh, 2017; Sothan, 2019; Vicario et al., 2018). In this case, researchers are less interested in explaining the phenomenon per se, while the focus is on finding the most performing model to forecast students who withdraw. In fact, the prediction of low performers is increasingly getting the attention of academics (Saa, Al-Emran, & Shaalan, 2019). In addition, data mining approach to education is fastly becoming an important field of research due to its ability to extract new knowledge about this aspect from a huge amount of students' data (Wook, Yusof, & Nazri, 2017).

The efforts behind the present study converge into the development of a clear theoretical framework, placed in midway between the two approaches, considering the educational process and the need of predicting students' outcome as early as possible. On one side, the environmental factors appeared not as external but as student's attributes; while, on the other side, the data-driven approach is substituted with an information-driven modelling. This study wants to go beyond traditional data-driven models; indeed, they focus attention on the algorithms' composition to improve prediction's performance, which depends on data typology and availability.

With the aim of filling the gaps within the two approaches, the theoretical framework proposed in this paper poses its basis on students' educational journey. This concept lays its foundation on Cunha and Heckman (2007), where the formation of individual skills (both cognitive and non-cognitive) is the result of a process where investments, environments and genes intervene. These factors interact and influence each other, to produce behaviours and abilities, overcoming the old "nature vs. nurture"

distinction. In fact, the individual characteristics are the results of both innate and acquired factors. The technology that governs this process is formed by sequential periods, which are *multistage* and *interrelated*, so each period is influenced by the previous one and, in turn, influences the next. This means that inputs and investments in each stage produce outputs, which will be inputs of next stages themselves. For the purpose of our framework, we consider the Cunha and Heckman (2007) educational stages as school cycles (see Figure 1): childhood, primary school, middle school and high school (we use "K12" to refer to all school's grades until the 12^{th}) and university.

Figure 1: The educational stages according to student's timeline.



During each stage, it is possible to gather different types of information about students. The collected information deal with educational path, such as grades or school data, or with personal and demographic information, for instance the citizenship or family's situation. It is highly relevant the moment in which information is collected: some demographic characteristics appears in the timeline at individuals' birth (for example, gender or date and place of birth), while examinations at middle schools are stored within the K12's group. The key feature of this model is that individual experiences enrich students' personal timeline.

Starting from the assumption that the process of skills' formation is *multistage* and *interrelated* (Cunha & Heckman, 2007), the milestone of the proposed framework relies on the possibility to predict student's dropout, considering groups of variables related to the educational stages, in different periods of time. In the perspective described above, educational data scientists may take into consideration how a single piece of information (e.g. single variable) may add an informative advantage to the prediction of dropout, in relation to the various educational stages. This approach allows the analysts to consider students' performance as the result of a cumulative process over time. Further and most important, educational data scientists may predict students' outcome (in this case dropout) standing on different points along the timeline. In other words, it is possible to predict student's outcome considering new variables or variables' group each time. This conception allows finding the optimal stage to observe student's outcome, facing the trade-off between prediction *accuracy*, which normally improve when adding more features, and the potential *timing to intervene*, that needs to be reduced as much as possible, so with early predictions. The proposed framework aims at addressing the managerial challenge for education: helping students deemed as at-risk the earliest moment possible.

2.2 The theoretical framework for Higher Education Institutions: practical application and academic literature

From an operational standpoint, a complete picture about students' career and personal characteristics cannot be obtained, so a *reduced* view of the proposed theoretical framework needs to contextualise it into real-world practice. Institutions have an incomplete outlook about student's educational path, mostly based on two types of variables: *dynamic*, such as the digital footprints students leave within their organization (Azcona, Hsiao, & Smeaton, 2019), and *static*, with demographic information usually registered at the enrolment moment and data about educational performance over time. Higher Edu-

cation Institutions (HEIs) observe students' academic performance and the associated characteristics, without knowing the broader picture of previous history. Universities can make hypotheses about their students' previous career, using available information. The practical action is that institutions and educational data scientists need to position themselves along the student's timeline and look at their present, past and available characteristics, to predict future outcome (for example, graduation *versus* dropout). Intuitively, the more information is available, the more accurate is the prediction. However, this is an optimization problem: from a managerial perspective, the timing of the prediction is equally important to its accuracy, e.g. an early prediction with 85 per cent of dropout prediction accuracy is preferable to a late one with 95 per cent of accuracy. The complete timeline from HEIs' perspective comprises students' information, grouped according to educational path stages, as illustrated in the previous paragraph: (i) demographic characteristics, (ii) previous studies information (K12 information) and (iii) academic performance. To confirm the choice of these groups of variables, Saa et al. (2019) develops an interesting meta-analysis taking into consideration 36 studies about dropout prediction at universities: it emerges that the determinants mostly related to dropout are previous grades and class performance, demographics, social information, instructor attributes, course attributes, course evaluations, environment, eLearning activity. From these results emerges that the selection of the three groups of features, as proposed in the current paper, is acceptable and generalisable.
From the university's standpoint, it is worth considering the academic performance group as the central one in the analysis. It can be divided into sub-groups, seeing academic career as a timeline too, composed by within-year periods (e.g. Winter and Spring terms). The various stages influence each other to contribute to the final outcome.

The reference theories concerning dropout started in 1970 when Spady (1970) associates the Durkheim theory of suicide (Friedman, 1952) with withdrawing. Breaking the ties with society, or university environment, means a lack of integration. The social interactions are the basis to be integrated into a community. Starting from Spady's work, many researchers retake this concept trying to figure out forces and features mostly explaining social integration (Cabrera & La Nasa, 2000; Cabrera et al., 1990; John et al., 1996; Pascarella & Terenzini, 1980; Tinto, 1975). These authors invest efforts in mapping and identifying the forces which are conducive to students' dropout. To confirm the choice of variables to be considered, most of these theories utilises demographic and academic information (i.e. family conditions, previous studies, experiences and academic performance) as inputs. Authors also map non-measurable factors to explain dropout, such as motivation, social integration and other non-cognitive features. The psychological aspects related to dropout decisions opens the discussion about the features which are too complicated and difficult to be integrated into a Machine Learning algorithm. Anyway, this interesting part requires separate discussions, combining different methodologies for its study.

When focusing only on the data-driven studies, the framework proposed is based on the academic articles, selected according to three conditions: (a) published after 2009, (b) apply Machine Learning techniques to predict dropout and (c) published on highly-ranked academic journal. The analysis of these academic papers focuses on highlighting the most powerful predictors of student's dropout. In particular, each study adopting Machine Learning algorithms, aside from showing algorithms' performance in terms of accuracy of predictions, displays individual determinants most correlated to dropout decision than institutional features. Based on this information, we aim at categorizing such determinants according to the variables' groups already defined in the theoretical framework. In particular, demographic information (it collects data about the individual, such as place and year of birth, gender, parents' info, etc.), previous studies (it collects information about previous schools, such as grades or school's information) and academic performance (it collects academic performance measurements, such

as exams, credits, course, etc.). For this reason, as shown in the Table 1, the cited studies are useful in justifying the adoption of the three variables' groups: both academic literature's streams, in fact, take into consideration, for the explanation of dropout phenomenon, a series of social and educational factors, which can be categorised into one (or more) of the three groups' variables. In this sense, Table 1 resumes the studies supporting each group in their prediction analysis. The proposed theoretical framework is useful for identifying factors associated with the risk of dropout, and as a consequence can represent a valuable help for decision-makers in setting remediation interventions to increase retention (something which is beyond the scope of the current paper, though). The present research helps to underline how the variables' groups may add a comprehensive advantage to this kind of modelling. In fact, this allows to take into consideration the educational history of the student, when predicting her future outcome.

Table 1: Variables' groups: their definitions and the supporting literature of data-driven studies.

	Demographic information	**Previous studies**	**Academic performance**
Definition	Personal characteristics	Information about previous schooling	Academic key performance indicator
Data-driven studies	(Belloc et al., 2010) (Seidel & Kutieleh, 2017) (Korhonen & Rautopuro, 2019) (Kotsiantis, Pierrakeas, & Pintelas, 2003) (Sothan, 2019) (Stratton, O'Toole, & Wetzel, 2008) (Arulampalam, Naylor, & Smith, 2004) (Caison, 2005) (Perez, Castellanos, & Correal, 2018) (Raju & Schumacker, 2015)	(Belloc et al., 2010) (Sothan, 2019) (Stratton et al., 2008) (Arulampalam et al., 2004) (Raju & Schumacker, 2015) (Seidel & Kutieleh, 2017) (Korhonen & Rautopuro, 2019) (Kotsiantis et al., 2003)	(Aulck, Velagapudi, Blumenstock, & West, 2016) (Stratton et al., 2008) (Caison, 2005) (Li et al., 2013) (Perez et al., 2018) (Khan, Al Sadiri, Ahmad, & Jabeur, 2019) (Raju & Schumacker, 2015)

Notes: the choice of each variables' group is supported by the listed literature. Data-driven studies, selected according to the date of publishing, the modelling's choice and the journal, are classified according to the predictors mostly related to dropout prediction which belong to one or more groups.

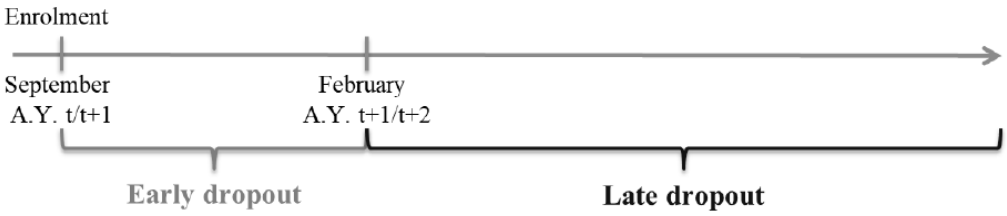
3 Materials and Methods

3.1 The context of Politecnico di Milano and technical details about its data

Politecnico di Milano (PoliMi) trains students in Engineering, Architecture and Design majors. It is now worldwide recognised as the most prestigious public university, positioning at the 1^{st} place according to QS ranking 2020. Worldwide, it is considered reference point for its research activity, appearing among the first 50 universities for Science, Technology, Engineering and Mathematics (STEM) disciplines. PoliMi counts $46,324$ enrolled students in Academic Year 2019/2020 in Bachelor and Master programs, among which $7,260$ are future architects, $4,305$ designers and $34,759$ engineers. On the teaching side, the university relies on $1,430$ professors, divided into the three disciplines - Engineering, Architecture and Design. In 2019, the Rector of the university creates a central team of Data Analytics for Institutional Support to improve decision-making process with an evidence-based approach. Extracting valuable information from students' learning behavior is one of team's main task, in the perspective of improving student life at PoliMi, defining new support activities for students and decreasing the number of students who dropout.

This study investigates the dimensions and characteristics of student dropout at PoliMi, making a further distinction between *early* and *late* dropout. In particular, early dropout occurs when the student drops within the 3^{rd} semester after enrolment, while late dropout when the student drops later on. For instance, let us suppose the student enrolled in September of the Academic Year 2009/2010, if she dropped before February 2011, it will be labelled as "early dropout", while if she dropped later as "late dropout" (see Figure 2).

Figure 2: Student's timeline to distinguish between early and late dropout.



From a methodological viewpoint, in an early warning system perspective, the group of variables about academic performance includes those until the end of first semester of the first year. Appendix A reports the list of variables used in the analysis with their explanation and descriptive statistics.

3.2 Methodology

The data we analyse cover students enrolled in different engineering programs at Politecnico di Milano. Students are nested within different degree programs, this induces a natural source of *dependence* among students enrolled in the same degree program. Therefore, the dependence structure among students is not homogeneous across the sample (i.e. it can be different across programs), but it has a latent structure, that is relevant and deserves to be taken into account. Classical regression models assume independence among observations and, therefore, they do not take into account this latent structure. On the other hand, multilevel regression models (Agresti, 2018; Goldstein, 2011; Pinheiro & Bates, 2006) are able to handle the hierarchical structure within the data and to model the structural dependence among them. These models disentangle the variability explained by each level of grouping of data and, therefore, help us in understanding the contribution given to each different aspect to the phenomenon of student dropout.

A second methodological aspect concerns the informative groups of variables as introduced in the theoretical framework. The academic literature indicates that the most powerful predictors for student dropout are categorised within three groups, that are (i) demographic information, (ii) previous studies and (iii) academic performance. These three areas of information are somehow sequential, in the sense that each group adds information, later in time, to the previous ones. Table 6 reports the list of variables we consider for each group. We are interested in measuring the importance of each group and, in the perspective of predicting student dropout as soon as possible, in identifying the minimum set - early in time - of information that allows us to have good predictions for student dropout. To this end, we start running our model considering only the first group and then adding the others sequentially, with the aim to see how the predictive power increases by adding each group of information and which is minimum amount of information we need to have an "accurate prediction". We start with the "poorest" model that considers only demographic information and we conclude with the most complete one, that

contains all the information (including early academic performance). In this way, we investigate both how much the predictive power increases by adding each group and which group registers the highest informative gain.

Lastly, generalised linear models are the most frequently used techniques in the literature to predict student dropout. Nonetheless, they impose a parametric functional dependence between the covariates and the response that sometimes may be too restrictive or unrealistic for data that describe complex contests (as the one we are exploring now). For this reason, we compare the results of generalised linear models with the ones obtained applying Machine Learning techniques, such as classification trees and Random Forest (Breiman, 2001; Hastie, Tibshirani, & Friedman, 2009). These are flexible methods able to investigate non linear associations among the covariates and the response and to model interactions among the covariates. Moreover, recent developments in this context allow classification trees to handle hierarchical data: in Fontana, Masci, Ieva, and Paganoni (2018) the authors propose a method to fit generalised mixed-effects regression trees (GMERT), while in Pellagatti, Masci, Ieva, and Paganoni (2020) the authors extend GMERT developing a new method to fit generalised mixed-effects random forest (GMERF). These methods have the strength and the flexibility of Machine Learning techniques and they still consider the nested structure of data.

In the light of this discussion, we run 18 different models, that are listed in Table 2.

Table 2: The different models for analysing early and late dropout prediction.

	Set of covariates included in the model					
	demographic info		demographic info + previous studies		demographic info + previous studies + academic performance	
generalised linear model	not nested	nested	not nested	nested	not nested	nested
classification tree	not nested	nested	not nested	nested	not nested	nested
random forest	not nested	nested	not nested	nested	not nested	nested

Our aim in specifying the different models is twofold: (i) the former is to find out which kind of models is able to well predict student dropout; (ii) the latter is, applying different models, to investigate how this predictive accuracy does change across models, how and when it increases and which kind of information we can derive from different model assumptions.

We recall now the basics of multilevel models, specifying their modelling both for the generalised linear models and for the tree-based methods. Let Y_{ij} be the binary variable that is equal to 1 if the $j-$th student within the $i-$th degree program, for $j = 1, \dots, n_i$ and $i = 1, \dots, N$, dropped his/her studies and equal to 0 otherwise. n_i is the total number of students who concluded their career (either dropped or graduated) enrolled in the $i-$th degree program and $N = 20$ is the total number of degree programs. Being Y_{ij} a Bernoulli variable where $Y_{ij} = 1$ with probability p_{ij} and $Y_{ij} = 0$ with probability $(1 - p_{ij})$, the classical logistic regression model (Agresti, 2018) takes the form:

$$\begin{aligned} &\mu_{ij} = \mathbb{E}[Y_{ij}] \qquad\qquad j = 1, \dots, n_i, \quad i = 1, \dots, N \\ &g(\mu_{ij}) = \eta_{ij} \\ &\eta_{ij} = \sum_{p=1}^{P+1} \beta_p x_{ijp} \end{aligned} \tag{1}$$

where $\mu_{ij} = p_{ij}$. p_{ij} is the probability that student j within degree program i drops, $g(\mu_{ij})$ is the logit link function, i.e. $g(\mu_{ij}) = logit(\mu_{ij}) = logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$, P is the total number of predictors, $\boldsymbol{\beta}$ is the $(P+1)-$dimensional vector of coefficients and $\mathbf{x}_{ij}$ is the $(P+1)-$dimensional vector of the covariates (including 1 for the intercept) relative to the (ij)-th observation. This modelling assumes that all the observations Y_{ij} (i.e. single students) are independent, that is to say, the production process of the outcome (dropout or not) is not affected by common factors across students.

If we now take into account the nested structure of data (i.e. students being enrolled into degree programs), the multilevel logistic regression model (Agresti, 2018), considering two levels, takes the following form:

$$\begin{aligned}
&\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{b}_i] \qquad\qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, N \\
&g(\mu_{ij}) = \eta_{ij} \\
&\eta_{ij} = \sum_{p=1}^{P+1} \beta_p x_{ijp} + \sum_{q=1}^{Q+1} b_{iq} z_{ijq} \\
&\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \Psi).
\end{aligned} \tag{2}$$

Conditionally on the random effects coefficients denoted by $\mathbf{b}_i$, the multilevel logistic regression model assumes that the elements of $\mathbf{Y}_i$ are independent. $\mathbf{z}_{ij}$ is the $(Q+1)-$dimensional vector of predictors for the random effects, $\mathbf{b}_i$ is the $(Q+1)-$dimensional vector of their coefficients and Ψ is the $(Q+1)\times(Q+1)$ within-group covariance matrix of the random effects. In multilevel models, fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters. In our case study, $\mathbf{b}_i$ are the coefficients relative to the $i-$th degree program.

Moving now to a Machine Learning (ML) approach, multilevel classification trees (Fontana et al., 2018) basically substitute the linear fixed-effects part in Eq. (2) with a classification tree structure:

$$\begin{aligned}
&\mu_{ij} = \mathbb{E}[Y_{ij}|\mathbf{b}_i] \qquad\qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, N \\
&g(\mu_{ij}) = \eta_{ij} \\
&\eta_{ij} = f(\mathbf{x}_{ij}) + \sum_{q=1}^{Q+1} b_{iq} z_{ijq} \\
&\mathbf{b_i} \sim \mathcal{N}(\mathbf{0}, \Psi)
\end{aligned} \tag{3}$$

where $f(\mathbf{x}_{ij})$ is not a linear combination of the coefficients $\boldsymbol{\beta}$ but it is a partition of the covariates space into boxes (or rectangles) and the prediction within each box is the mode of all the observations that belong to that box. The absence of a specific functional form makes this method very flexible and able to better model interactions among the covariates. Similarly, multilevel random forest (Pellagatti et al., 2020) takes the form in Eq. (3), where $f(\mathbf{x}_{ij})$, instead of being a standard classification tree, is a random forest, that is an ensemble of classification trees. Random forest basically works taking many training sets from the entire population, building a separate prediction model using each training set, and averaging the resulting predictions. Moreover, during this process, it considers different subsets of covariates for each training set, in order to give all variables the possibility to be taken into account in the tree splits - avoiding the risk that some variables cover the effect of other less significant ones

(Hastie et al., 2009). Therefore, the advantage of random forest is twofold: it reduces the model variance and it handles the presence of highly correlated covariates, disentangling their associations with the response variable. Random forest gives as output the importance ranking of the covariates in predicting the response, measured as the mean decrease in Gini index - i.e. we can add up the total amount that the Gini index is decreased by splits over a given predictor, averaged over all trees of the ensemble (Raileanu & Stoffel, 2004).

4 Results

The sample of students with ended career that we consider is composed by three categories of students: (i) graduated students, (ii) students who early dropped out (within the first three semesters after the enrolment), (iii) students who late dropped out (after the first three semesters after the enrolment). Since our aim is to investigate the determinants of both the two types of dropout, we run the models in Table 2 twice: one considering *graduated students versus early dropout students*, and the other considering *graduated students versus late dropout students*.
We train our models on a training set, that is composed by students with ended career enrolled between a.y. 2010/2011 and a.y. 2014/2015 and we test it on a test set composed by students with ended careers enrolled in a.y. 2015/2016.

In particular, in our models, $Y_{ij} = 1$ when student j within degree program i dropped, early or late depending on the model setting, and $Y_{ij} = 0$ when he or she graduated; $\mathbf{X}$ is the matrix of the fixed-effects covariates that contain all student-level characteristics shown in Table 6 (sequentially included in the models, according to the different groups) and, lastly, when running multilevel models, i.e. when we take into account the hierarchical structure of students nested within degree programs, we include in the random effects part only a random intercept, i.e.

$$\begin{aligned} &\mu_{ij} = \mathbb{E}[Y_{ij}|b_i] \qquad\qquad j = 1, \dots, n_i, \quad i = 1, \dots, N \\ &g(\mu_{ij}) = \eta_{ij} \\ &\eta_{ij} = f(\mathbf{x}_{ij}) + b_i \\ &b_i \sim \mathcal{N}(0, \sigma^2_\psi) \end{aligned} \tag{4}$$

where b_i is the value-added given by the $i-$th degree program to the dropout probability: if b_i is negative, students within the $i-$th degree program are on average less likely to drop with respect to the others; while, if b_i is positive, students within the $i-$th degree program are on average more likely to drop with respect to the others.

Given Eq. (4), $f(\mathbf{x}_{ij})$ is equal to a linear combination of the fixed-effects covariates in the case of a multilevel linear model, to a classification tree in the case of a multilevel classification tree and to a random forest in the case of a multilevel random forest.

In order to compare the performance of the fitted models, we compute two indexes: (i) the Area Under the ROC Curve (AUC), that provides an aggregate measure of performance across all possible classification thresholds; (ii) the sensitivity index. We choose to measure the sensitivity index among the set of possible performance indexes because we are interested in finding the model that better identifies the students at risk, i.e., the model with highest sensitivity. In Tables 3 and 4, we report

the results of the fitted models in terms of AUC and sensitivity, for early and late dropout prediction, respectively.

Table 3: Area Under the Curve (AUC) and sensitivity index (sens) of the 18 models run for early dropout *versus* graduated.

	Set of covariates included in the model					
	demographic info		demographic info + previous studies		demographic info + previous studies + academic performance	
generalised linear model	**not nested** AUC: 0.568 sens: 0.099	**nested** AUC: 0.593 sens: 0.312	**not nested** AUC: 0.626 sens: 0.465	**nested** AUC: 0.641 sens: 0.506	**not nested** AUC: 0.971 sens: 0.888	**nested** ***AUC: 0.972*** ***sens: 0.889***
classification tree	**not nested** AUC: 0.532 sens: 0.083	**nested** AUC: 0.581 sens: 0.267	**not nested** AUC: 0.533 sens: 0.075	**nested** AUC: 0.621 sens: 0.507	**not nested** AUC: 0.887 sens: 0.778	**nested** AUC: 0.946 sens: 0.888
random forest	**not nested** AUC: 0.528 sens: 0.006	**nested** AUC: 0.583 sens: 0.291	**not nested** AUC: 0.586 sens: 0.101	**nested** AUC: 0.639 sens: 0.415	**not nested** AUC: 0.967 sens: 0.870	**nested** ***AUC: 0.968*** ***sens: 0.871***

Notes: The sensitivity is obtained as sensitivity $= \frac{\#\ true\ positive}{\#\ true\ positive + \#\ false\ negatives}$*, where the true positives are the students correctly classified as dropout by the model and the false negatives are the students that are wrongly identified as graduated by the model.*

Table 4: Area Under the Curve (AUC) and sensitivity index (sens) of the 18 models run for late dropout *versus* graduated.

	Set of covariates included in the model					
	demographic info		demographic info + previous studies		demographic info + previous studies + academic performance	
generalised linear model	**not nested** AUC: 0.689 sens: 0.403	**nested** AUC: 0.719 sens: 0.443	**not nested** AUC: 0.768 sens: 0.566	**nested** AUC: 0.783 sens: 0.601	**not nested** AUC: 0.956 sens: 0.822	**nested** ***AUC: 0.957*** ***sens: 0.824***
classification tree	**not nested** AUC: 0.662 sens: 0.261	**nested** AUC: 0.696 sens: 0.379	**not nested** AUC: 0.620 sens: 0.261	**nested** AUC: 0.751 sens: 0.5526	**not nested** AUC: 0.869 sens: 0.752	**nested** AUC: 0.946 sens: 0.831
random forest	**not nested** AUC: 0.643 sens: 0.263	**nested** AUC: 0.707 sens: 0.388	**not nested** AUC: 0.719 sens: 0.324	**nested** AUC: 0.777 sens: 0.548	**not nested** AUC: 0.939 sens: 0.774	**nested** ***AUC: 0.948*** ***sens: 0.778***

Notes: The sensitivity is obtained as sensitivity $= \frac{\#\ true\ positive}{\#\ true\ positive + \#\ false\ negatives}$*, where the true positives are the students correctly classified as dropout by the model and the false negatives are the students that are wrongly identified as graduated by the model.*

Multilevel *versus* Classical models

From Tables 3 and 4, we can highlight that the predictive performance, both in terms of AUC and sensitivity, is always higher in multilevel models than the one in the correspondent models that do not take into account the nested nature of data, all else equal. This finding suggests that taking into account the nested structure of students within degree programs improves the performance of the model and identifies a source of variability within students performance that is due to their grouping structure. The difference between the two types of models - i.e. multilevel and not multilevel - is particularly relevant when we consider only demographic information or demographic information and previous studies as covariates while it decreases when we consider also the academic performance. Hereafter, we focus our attention on the results of multilevel (nested) models.

The importance of student-level covariates

Considering now the three groups of covariates, we observe that the predictive performance of models when we consider only demographic information of students is quite low and it does not increase much when we add previous studies among the covariates. Instead, we observe a sharp increase in predictive power when we add information about the academic performance, reaching very high predictive performance. This result suggests that the background of students alone is not sufficient to obtain a good proxy of their academic career, while a very big portion of variability in students dropout is explained by students academic performance during the first semester. This evidence points at confirming the importance to monitor the early performance of students as a good indicator for their subsequent results.

Parametric models *versus* Machine Learning

From the comparison of generalised linear models with classification trees and random forest, it emerges that classification trees have almost always lower predictive power than generalised linear models and random forest. In particular, the performance of multilevel generalised linear models and multilevel random forest are very close to each other and are the best ones. We therefore focus on the interpretation of multilevel generalised linear models and multilevel random forest results, to compare them and to investigate which kind of insights about student dropout phenomenon we can extract from these two different models.

4.1 Multilevel generalised linear model *versus* Multilevel random forest

Table 5 reports the results of the multilevel generalised linear models, adding the three groups of covariates sequentially, both for early and late dropout, respectively.

Table 5: Results of multilevel linear models in Eq. (4) for early and late dropout respectively, considering as fixed-effects covariates: (1) demographic information, (2) demographic information and previous studies, (3) demographic information, previous studies and academic performance in the first semester.

	Dependent variable:					
	Early dropout *vs* graduated			Late dropout *vs* graduated		
	(1e)	(2e)	(3e)	(1l)	(2l)	(3l)
Constant	-5.446^{***}	-1.703^{***}	1.499^{**}	-9.866^{***}	-4.866^{***}	-2.418^{***}
	(0.338)	(0.369)	(0.585)	(0.378)	(0.399)	(0.469)
Gender = Male	0.058	0.173^{***}	0.207^{**}	0.700^{***}	0.799^{***}	0.594^{***}
	(0.047)	(0.049)	(0.088)	(0.062)	(0.064)	(0.076)
Student's origin =	-0.028	0.030	0.281^{***}	-0.178^{***}	-0.128^{***}	0.068
Native Out of Milan	(0.043)	(0.044)	(0.076)	(0.045)	(0.047)	(0.058)
Student's origin =	0.873^{***}	0.256	-0.208	1.371^{***}	0.697^{***}	0.686^{**}
Non-Italian abroad	(0.228)	(0.261)	(0.474)	(0.202)	(0.242)	(0.311)
Student's origin =	0.164	-0.189	0.187	0.848^{***}	0.487^{***}	0.537^{***}
Non-Italian out of Milan	(0.179)	(0.191)	(0.331)	(0.142)	(0.155)	(0.197)
Student's origin =	0.352^{**}	-0.351^{*}	-0.031	1.300^{***}	0.631^{***}	0.477^{**}
Non-Italian in Milan	(0.169)	(0.193)	(0.331)	(0.134)	(0.153)	(0.197)
Access To Studies Age	0.207^{***}	0.156^{***}	-0.033	0.411^{***}	0.342^{***}	0.168^{***}
	(0.018)	(0.018)	(0.026)	(0.019)	(0.019)	(0.021)
Previous School =		0.302^{***}	0.120		-0.028	-0.181
Classic		(0.072)	(0.126)		(0.094)	(0.114)
Previous School =		0.526^{***}	0.214		0.472^{***}	0.174
Other		(0.108)	(0.191)		(0.114)	(0.142)
Previous School =		0.019	-0.058		0.138^{**}	0.230^{***}
Technical		(0.059)	(0.100)		(0.056)	(0.069)
Admission Score		-0.041^{***}	0.012^{***}		-0.054^{***}	-0.005^{**}
		(0.002)	(0.003)		(0.002)	(0.003)
Total Credits 1s			-0.223^{***}			-0.163^{***}
			(0.004)			(0.003)
Attempts 1s >1			-0.775^{***}			0.361^{***}
			(0.088)			(0.078)

	(1e)	(2e)	(3e)	(1l)	(2l)	(3l)
Attempts 1s = 0			2.368***			1.161***
			(0.250)			(0.256)
Change Degree			0.084			0.037
			(0.086)			(0.066)
Income= DSU			−0.325			−0.809***
			(0.282)			(0.233)
Income= high			−0.068			−0.097
			(0.086)			(0.068)
Income= low			0.003			0.137**
			(0.089)			(0.068)
Income= unknown			−1.337			−0.816*
			(1.024)			(0.450)
Observations	19,803	19,803	19,803	19,660	19,660	19,660
Log Likelihood	−9,202.492	−8,867.000	−3,354.970	−8,182.608	−7,691.066	−5,297.633
Akaike Inf. Crit.	18,420.980	17,758.000	6,749.940	16,381.220	15,406.130	10,635.270
Bayesian Inf. Crit.	18,484.130	17,852.720	6,907.812	16,444.310	15,500.770	10,792.990

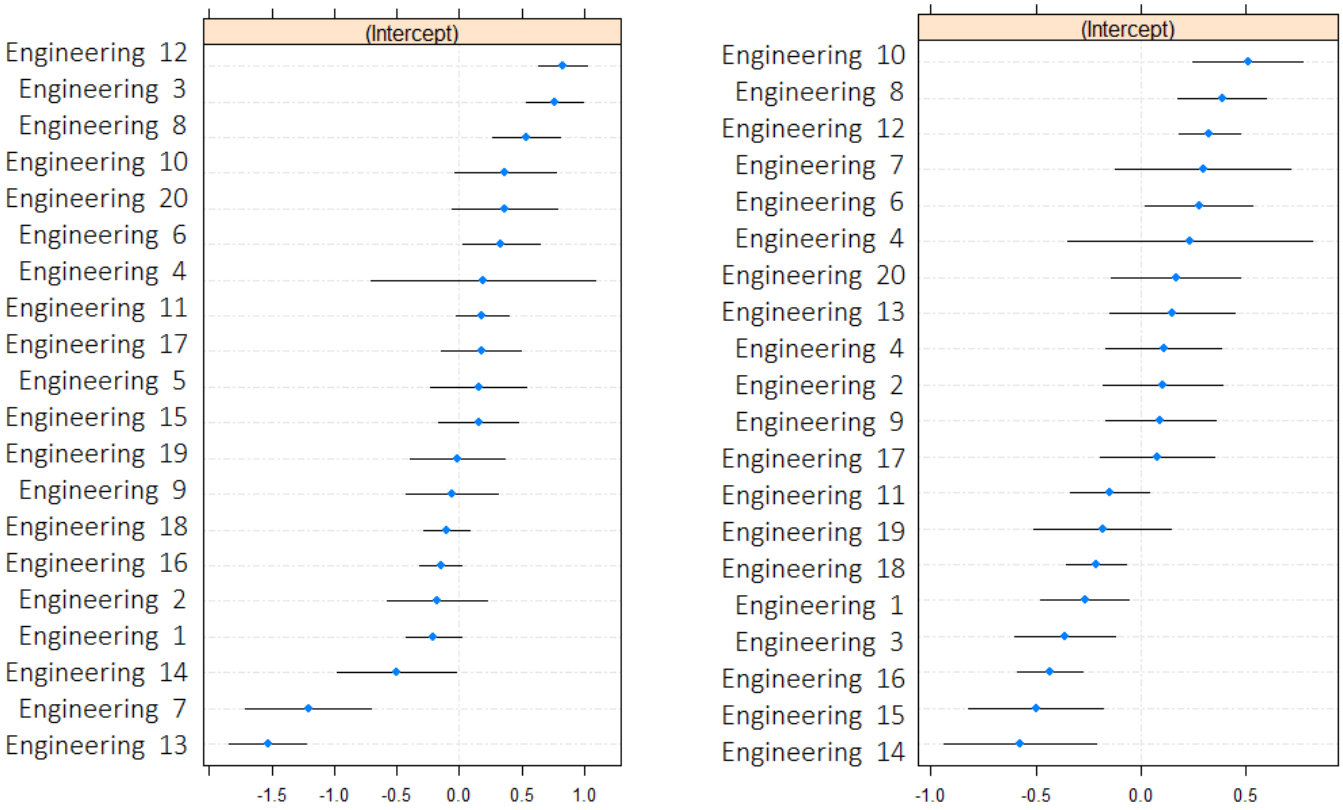
*Notes: Results are reported in terms of regression coefficients point estimates with their standard deviation (in brackets) Stars represent the statistical significance: *p<0.1; **p<0.05; ***p<0.01.*

We observe that, by adding groups of covariates, the importance of the previous group changes, both for early and late dropout predictions. Since we believe that is essential to take into account at least the first semester career of students in order to have an accurate proxy of their dropout probability, we directly look at the complete models, the ones with highest predictive performances, for early and late dropout, i.e. models (3e) and (3l), that include all the three groups of covariates. In this way, we can interpret the net associations of the demographics and the previous studies with the dropout probability, after adjusting for the first semester career. By looking at models (3e) and (3l) coefficients, several interesting observations emerge: males are more likely to early and especially late drop than females; Native Italians off-site are more likely to early drop than Native Italians in-site, while they do not differ in terms of late dropout probability; non-Italian students, either in-site, off-site or not having residence in Italy, are more likely to late drop than Native Italians in-site; students starting their careers at PoliMi at an older age than the average, are more likely to late drop; students who attended technical high schools are more likely to late drop than the ones who attended scientific high schools; the higher is the admission test score at PoliMi, the higher is the probability of students early dropout and the lower is the probability of students late dropout; the higher is the number of credits obtained at the first semester, the lower are both the early and late dropout probabilities; students doing more than one attempts per exam during the first semester are less likely to early drop and more

likely to late drop with respect to students doing one attempt per exam; students that do not attempt any exam during the first semester are more likely both to early and late drop with respect to students doing one attempt per exam; if students change degree program during their career does not seem to be significant; lastly, regarding the students family income, it does not result to be significantly associated to early dropout probability, while students with DSU, low income or unknown income are more likely to late drop than students with maximum income group.

Regarding the random effects, the estimated random intercepts, $\hat{b}_i$, for $i = 1, \ldots, N$, that represent the value-added, either positive or negative, of the 20 degree programs to the dropout probability of their students are reported in Figure 3, together with their confidence intervals. Degree programs with estimated $\hat{b}_i$ whose confidence interval is totally positive (or negative) have on average students more (or less) likely to dropout, all else equal.

Figure 3: Estimated random intercepts, $\hat{b}_i$, for $i = 1, \ldots, N$, with their confidence intervals, of the 20 degree programs estimated in Eq. (4) for the generalised linear model, for early dropout (panel (a)) and late dropout (panel (b)).



In order to measure the magnitude of the random effects, we compute the Variance Partitioning Coefficient (VPC) (Goldstein, Browne, & Rasbash, 2002) that represents the percentage of unexplained

variability in the response that is given to the grouping level (degree programs). In particular:

$$VPC_{early} = \frac{\sigma^2_{\psi_{early}}}{\sigma^2_{\psi_{early}} + \pi^2/3} = 9.6\% \qquad \text{and} \qquad VPC_{late} = \frac{\sigma^2_{\psi_{late}}}{\sigma^2_{\psi_{late}} + \pi^2/3} = 8.1\%, \tag{5}$$

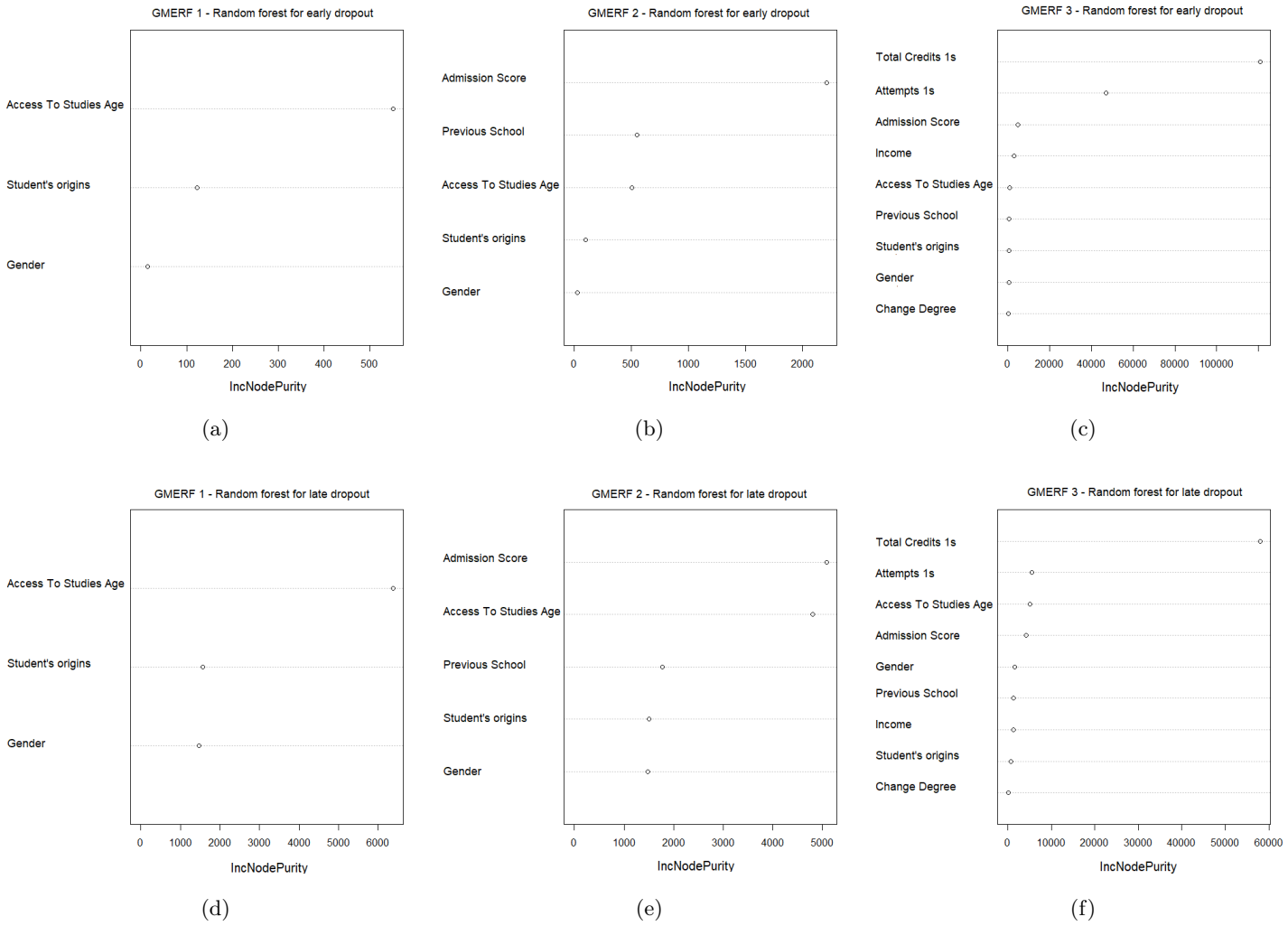
meaning that 9.6 per cent of the unexplained variability in student early dropout phenomenon is given by the grouping of students within degree programs and, equivalently, 8.1 per cent of the unexplained variability in student late dropout is given by the grouping of students within degree programs. These percentages reach almost 10 per cent of student variability, suggesting that for equal student characteristics there is still heterogeneity in the likelihood of student dropout across degree programs, as there are certain degree programs in which students are more likely to leave their studies than in others, or vice-versa.

By looking at Figure 3 and considering the VPCs, we observe that the magnitude of the degree program effects on early and late dropout is similar, but still a bit higher on the early dropout phenomenon. Degree programs that have significantly different from zero effect, vary between early and late dropout prediction. In particular, there are degree programs whose effect is coherent between early and late dropout: Students in Engineering 8, 10 and 12 are more likely to both early and late drop with respect to other students; Students in Engineering 1, 14 and 16 are less likely to both early and late drop. On the opposite, there are degree programs whose effect is very different between early and late dropout: students in Engineering 3, for example, are more likely to early drop but less likely to late drop with respect to the average; students in Engineering 13 are in line with the average regarding the late dropout but they are less likely to early drop. Differences among degree programs might be due to various aspects: internal difficulties of degree programs, structural differences or movement of students from certain engineering courses to other faculties due to external drivers. With the available data it is not possible to investigate these mechanisms more profoundly, but we will explore this topic in future research.

We then focus on the results of multilevel random forest. As for the linear case, we extract information about both the fixed-effects part and the estimates of the random intercepts. Regarding the fixed-effects part, random forest give us as output the importance ranking of the covariates in predicting student dropout, measured as the mean decrease in Gini index. Figure 4 reports the variable importance plots computed by the random forest in Eq. (4) with the 3 groups of covariates added sequentially, both for early and late dropout. By looking both at the top and bottom panels of Figure 4, we observe that the importance of the covariates of the first two groups is extremely small when compared to the one of the covariates in the third group, both for early and late dropout. Indeed, when considering the entire set of covariates (panels c and f), the most important covariates are the ones regarding the first semester career at the university. In particular, for early dropout, the most important covariates are the number of total credits obtained and the average attempts per exam, while, for late dropout, the number of total credits explains, alone, almost all the variability in the response. Indeed, the predictive powers of *TotalCredits1sem* and *Attempts1sem* have a different order of magnitude with respect to the other covariates. It is reasonable to think that the average number of attempts varies more between the ones who drop immediately and the one who graduate than between the ones who drop after more than one year and the one who graduate. Indeed, it is likely possible that who drops immediately does neither attempt any exams, while who drops after one year, attempts more times without succeeding, before to dropout.

Regarding the random-effects part, Figure 5 reports the random intercepts estimated for each degree program, together with their confidence intervals. Comparing Figures 3 and 5, we observe that the degree course intercepts estimated by the multilevel generalised linear model and the multilevel

Figure 4: Variable importance plots computed by the random forest fixed-effects part in Eq. (4), adding the three groups of covariates sequentially for both early and late dropout prediction.

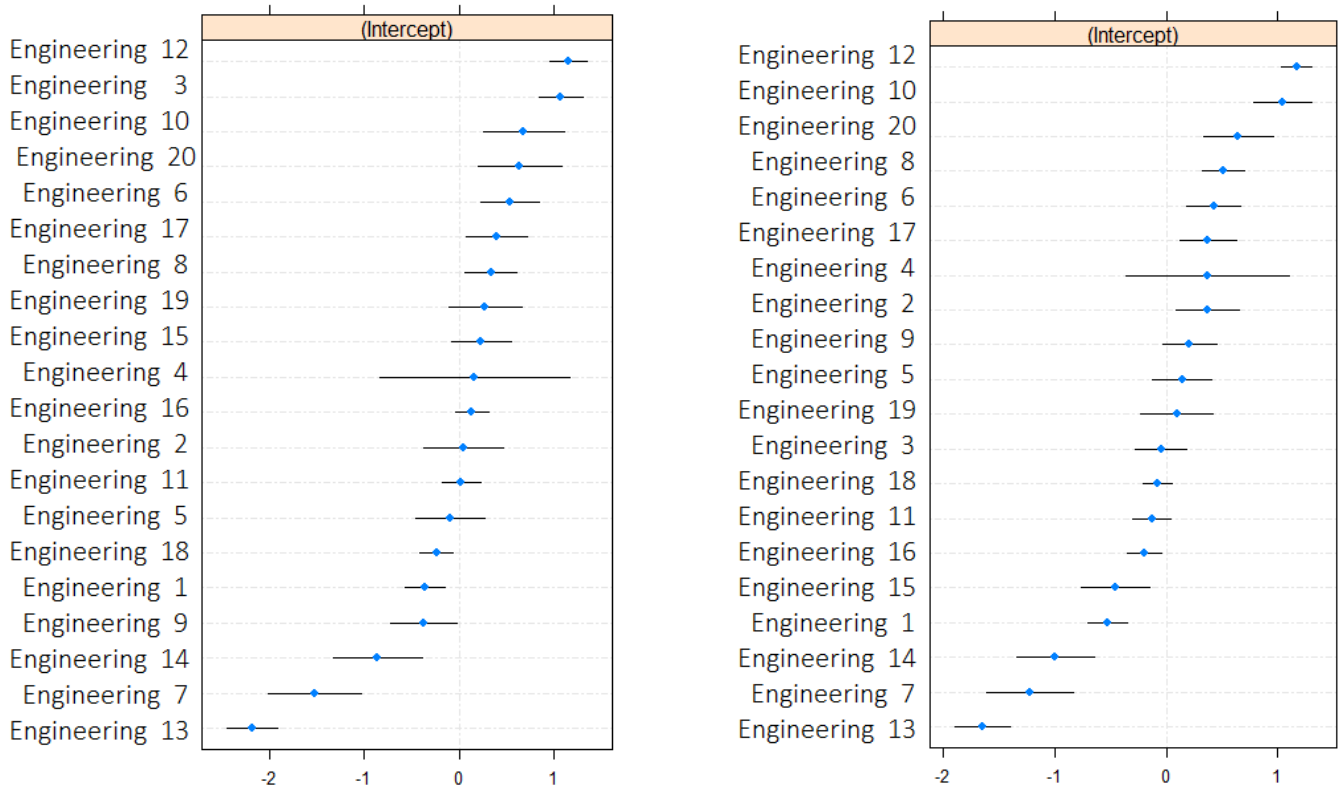


Notes: panels (a), (b) and (c) regard the models for predicting early dropout considering as covariates demographic information, demographic information + previous studies and demographic information + previous studies + academic performance, respectively; panels (d), (e) and (f) regard the models for predicting late dropout considering as covariates demographic information, demographic information + previous studies and demographic information + previous studies + academic performance, respectively.

random forest are coherent. In particular, when considering the early dropout phenomenon, the degree courses that have significant positive or negative effects identified by the two models are the same, while, regarding the late dropout phenomenon, they are mostly the same apart from Engineering 7

and 13 whose estimates of multilevel random forest are negative, while and the ones of multilevel generalised linear model are not statistically different from zero.

Figure 5: Estimated random intercepts, $\hat{b}_i$, for $i = 1, \ldots, N$, with their confidence intervals, of the 20 degree programs estimated in Eq. (4) for the random forest case, for (a) early dropout and (b) late dropout.



The VPCs estimated by multilevel random forest models are:

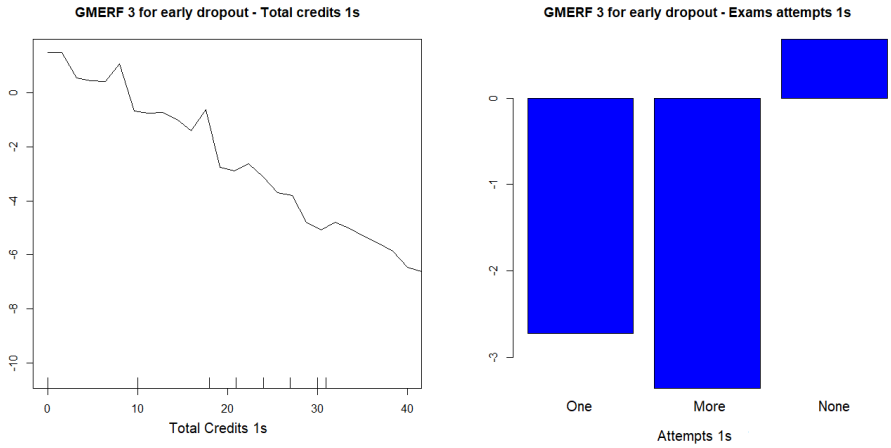
$$VPC_{early} = \frac{\sigma^2_{\psi_{early}}}{\sigma^2_{\psi_{early}} + \pi^2/3} = 16.1\% \qquad \text{and} \qquad VPC_{late} = \frac{\sigma^2_{\psi_{late}}}{\sigma^2_{\psi_{late}} + \pi^2/3} = 13.3\%, \tag{6}$$

that are slightly higher than the ones in Eq. (5), obtained by multilevel generalised linear models. It can be the case that random forests are better able to identify the true effect of different programs on the dropout phenomenon. In all the models, the VPCs estimated for early dropout response are higher then the VPCs estimated for late dropout response, suggesting that there is more heterogeneity in student early dropout across degree programs than the heterogeneity in student late dropout.

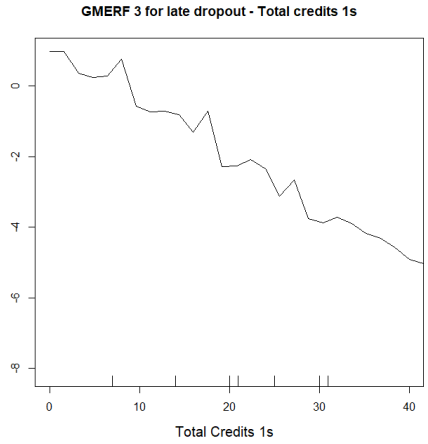
Variable importance plots showed in Figure 4 identify which are the most important variables in predicting student dropout, but they do not give us information about the type of associations (either

direct or inverse) of these variables with the response. To this end, partial dependence plots (Breiman, 2001) show the partial association between the response variable (student dropout probability) and each covariate, net to the effects of all the other covariates. Taking into account panels (c) and (f) of Figure 4, we select the most important variables in each of these two panels and we show in Figure 6 their partial dependence plots. These plots are particularly useful for exploring non-linear associations between the variables of interest.

Figure 6: Partial dependence plots of the most important variables (extracted from Figure 4) in predicting early and late student dropout.



(a) (b)



(c)

Notes: The X-axis reports the range of the covariate, while the Y-axis report the change in the predictor η_{ij} (that is directly proportional to the dropout probability p_{ij}) relative to the considered covariate. Panels (a) and (b) show the partial plots of Total Credits 1sem *and* Attempts 1sem*, respectively, relative to the probability of student early dropout; panel (c) shows the partial dependence plot of* Total Credits 1sem *relative to the probability of student late dropout.*

By looking at panel (b) of Figure 6, we see that doing zero attempts per exam, is associated to higher early dropout probability, while doing one and especially more than one attempts per exam is associated to lower early dropout probability. By looking at panels (a) and (c), we see that the probability of both early and late student dropout decreases when the number of credits obtained

at the first semester increases and, moreover, the number of credits is linear with the predictor η. This evident linear association is confirmed also by the very low p-value that this covariate has in the multilevel generalised linear model and explains the reason why the performances of multilevel generalised linear models are at the same level or even better than the ones of multilevel random forest. Indeed, being *Total Credits 1sem* the most significant covariate, i.e. the one with the highest predictive power, and having it a linear association with the predictor η, generalised linear models perform better than random forest. This evidence explains also the higher VPCs of multilevel random forest with respect to the ones of multilevel generalised linear models. Indeed, given that the tree-based structure worse fit the fixed-effects part (whose most important characteristic is the linearity of the *Total Credits 1sem* effect), the percentage of unexplained variability at student level is higher and the random effects part gains power.

This specific finding raises a final reflection about the use of ML techniques. They are known to be very flexible and to perform good prediction results in complex data structures, when non-linearities and interactions are at play. In our case, the situation is partly different. Here, the dropout phenomenon is characterised by some important linear relationships. Moreover, the theory helps us in identifying some important variables a priori. In these circumstances, the parametric models (well trained to fit the unknown functional form) are actually able to obtain good predictions.

Appendix B reports a reflection about the robustness of model choices. In particular, we run a sensitivity analysis comparing the performances of the model when considering first semester *versus* first year (first two semesters) student career information. Results confirm that the early warning system (first semester information model) works sufficiently well: the gain in prediction performance when considering the entire first year is not so high to justify the waiting until the end of the first year to identify the students at risk.

5 Concluding remarks

This paper demonstrates how useful and powerful can data analysis be, which estimates the likelihood that a student drops out in the first year of university attendance, or even later. Providing decision-makers with adequate tools for such predictions is a critical development of HE management nowadays. If properly used, the quantitative evaluation of factors associated with higher dropout risk can help reducing the loss of human capital, by retaining more students. Moreover, these systems hold the promise of improving the efficiency and effectiveness of universities' operations. In such a perspective, a functioning Early Warning System (EWS) should become part of the toolbox of any HE institution in the next years.

University managers and decision-makers should become familiar with data produced by EWSs. This is a new core competence, that can be acquired and/or strengthened with some formative interventions. The findings presented in this research advance the state-of-knowledge in this field. Specifically, the main results derived from the empirical analysis can be summarised in three key messages.

First, the different algorithms developed and tested in developing the empirical (final) model perform very similarly. This evidence should reassure the analysts and decision makers about the robustness of the findings - with the consequence that the algorithms can be used immediately. Second, the inclusion of academic results at the end of the first semester dramatically improves the quality of predictive modelling of dropout. Interestingly, these variables are much more predictive and relevant than demographic characteristics and previous achievement during high school (this finding is consistent with (Von Hippel & Hofflinger, 2020)). Third, the consideration of the grouping variable (i.e. the degree program chosen by each student) also plays a central role. Indeed, the outcomes of

the empirical analysis clearly reveals heterogeneous incidence of attending a specific degree program. Thus, a multilevel structure of both linear models and ML techniques has been specifically set and employed for developing the final version of algorithms to be used.

The critical reading of these messages, coupled with the interpretation of the theoretical framework that we propose in Section 2, suggests two major practical implications of the present study. (i) On one side, the amount and quality of available information is a notable condition for the EWS to work properly. The database built with PoliMi data is enough broad and complete - indeed, the prediction of at-risk students works pretty well. At the same time, the collection of data might be expanded in scope for taking individuals' beliefs, motivation and attitudes into account. These soft elements are central for maximizing students' performance, but they are not monitored regularly and in a structured way. A clear suggestion is to create surveys and automated collection methods for asking students more details about perceptions of their own internal life. Self-reported information can be important here. (ii) On the other side, the systematic adoption of an online platform and digital evaluation systems can further improve the institution' s ability of monitoring at-risk students. To the extent that teaching activities are weekly supported by digital platforms, the (dynamic) analysis of students' performance over time can be added to the model for detecting poor-performers and at-risk students early on time.

If the findings presented in this paper are accepted as credible and relevant, the way forward is constituted by two steps. The first consists of the establishment of a Unit dedicated to Data Analytics at the level of the single institution. Building capacity in each university is useful, and creating internal units for "learning from its data" is a necessary development of managerial responsibilities. The second step deals with the design and implementation of interventions for supporting at-risk students. The literature highlights and suggests different experiences, which validity and replicability must be assessed carefully. Also, each intervention must be accompanied by a rigorous evaluation procedure, aiming at understanding "what works" and weaknesses. Overall, experimental design can be favored when possible. Politecnico di Milano undertakes this approach, and we hope to validate soon a rigorous protocol for remedial interventions, as well as for the evaluation of their impact.

Acknowledgement

We are grateful to the IT Office of PoliMi for their support in extracting data and pre-processing them.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Aina, C. (2013). Parental background and university dropout in italy. *Higher Education*, *65*(4), 437–456.

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. *Higher Education Studies*, *6*(2), 1–18.

Anvur: Rapporto biennale sullo stato del sistema universitario e della ricerca. (2018). (https://www.anvur.it/rapporto-biennale/rapporto-biennale-2018)

Arulampalam, W., Naylor, R., & Smith, J. (2004). Factors affecting the probability of first year medical student dropout in the uk: a logistic analysis for the intake cohorts of 1980–92. *Medical Education*, *38*(5), 492–503.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *arXiv preprint arXiv:1606.06364*.

Azcona, D., Hsiao, I.-H., & Smeaton, A. F. (2019). Detecting students-at-risk in computer programming classes with learning analytics from students' digital footprints. *User Modeling and User-Adapted Interaction*, *29*(4), 759–788.

Barbu, M., Vilanova, R., Vicario, J., Pereira, M. J., Alves, P., Podpora, M., ... others (2019). Data mining tool for academic data exploitation: Publication report on engineering students profiles. *ERASMUS+ KA2/KA203*.

Belloc, F., Maruotti, A., & Petrella, L. (2010). University drop-out: an italian experience. *Higher education*, *60*(2), 127–138.

Belloc, F., Maruotti, A., & Petrella, L. (2011). How individual characteristics affect university students drop-out: a semiparametric mixed-effects model for an italian case study. *Journal of Applied Statistics*, *38*(10), 2225–2239.

Bratti, M., Checchi, D., & De Blasio, G. (2008). Does the expansion of higher education increase the equality of educational opportunities? evidence from italy. *Labour*, *22*, 53–88.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brunori, P., Peragine, V., & Serlenga, L. (2012). Fairness in education: The italian university before and after the reform. *Economics of Education Review*, *31*(5), 764–777.

Burgos, C., Campanario, M. L., de la Peña, D., Lara, J. A., Lizcano, D., & Martínez, M. A. (2018). Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering*, *66*, 541–556.

Cabrera, A. F., & La Nasa, S. M. (2000). Understanding the college-choice process. *New directions for institutional research*, *2000*(107), 5–22.

Cabrera, A. F., Stampen, J. O., & Hansen, W. L. (1990). Exploring the effects of ability to pay on persistence in college. *The Review of Higher Education*, *13*(3), 303–336.

Caison, A. L. (2005). Determinants of systemic retention: Implications for improving retention practice in higher education. *Journal of College Student Retention: Research, Theory & Practice*, *6*(4), 425–441.

Contini, D., Cugnata, F., & Scagni, A. (2018). Social selection in higher education. enrolment, dropout and timely degree attainment in italy. *Higher Education*, *75*(5), 785–808.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Daniel, B. (2015). B ig d ata and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, *46*(5), 904–920.

De Freitas, S., Gibson, D., Du Plessis, C., Halloran, P., Williams, E., Ambrose, M., ... Arnab, S. (2015). Foundations of dynamic learning analytics: Using university student data to increase retention. *British journal of educational technology*, *46*(6), 1175–1188.

Di Pietro, G., & Cutillo, A. (2008). Degree flexibility and university drop-out: The italian experience. *Economics of Education Review*, *27*(5), 546–555.

Fontana, L., Masci, C., Ieva, F., & Paganoni, A. M. (2018). *Performing learning analytics via generalized mixed-effects trees.* MOX-report n. 43/2018.

Friedman, P. (1952). Suicide: By emile durkheim. translated by john a. spaulding and george simpson. edited with an introduction by george simpson. *Psychoanalytic Quarterly*, *21*, 416–419.

Ghignoni, E. (2017). Family background and university dropouts during the crisis: the case of italy. *Higher Education*, *73*(1), 127–151.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *1*(4), 223–231.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

John, E. P. S., Paulsen, M. B., & Starkey, J. B. (1996). The nexus between college choice and persistence. *Research in Higher Education*, *37*(2), 175–220.

Khan, I., Al Sadiri, A., Ahmad, A. R., & Jabeur, N. (2019). Tracking student performance in introductory programming by means of machine learning. In *2019 4th mec international conference on big data and smart city (icbdsc)* (pp. 1–6).

Korhonen, V., & Rautopuro, J. (2019). Identifying problematic study progression and "at-risk" students in higher education in finland. *Scandinavian Journal of Educational Research*, *63*(7), 1056–1069.

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *International conference on knowledge-based and intelligent information and engineering systems* (pp. 267–274).

Larrabee Sønderlund, A., Hughes, E., & Smith, J. (2019). The efficacy of learning analytics interventions in higher education: A systematic review. *British Journal of Educational Technology*, *50*(5), 2594–2618.

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education - a literature review. In *Learning analytics: Fundaments, applications, and trends* (pp. 1–23). Springer.

Li, K. F., Rusk, D., & Song, F. (2013). Predicting student academic performance. In *2013 seventh international conference on complex, intelligent, and software intensive systems* (pp. 27–33).

OECD. (2019). Education at a glance 2019: Oecd indicators. (https://doi.org/10.1787/f8d7880d-en)

Oppedisano, V. (2011). The (adverse) effects of expanding higher education: Evidence from italy. *Economics of Education Review*, *30*(5), 997–1008.

Pascarella, E. T., & Terenzini, P. T. (1980). Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The journal of higher education*, *51*(1), 60–75.

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2020). *Generalized mixed effects random forest: a flexible application to predict university student dropout.* MOX-report, n. 36/2020.

Perez, B., Castellanos, C., & Correal, D. (2018). Applying data mining techniques to predict student dropout: a case study. In *2018 ieee 1st colombian conference on applications in computational intelligence (colcaci)* (pp. 1–6).

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in s and s-plus.* Springer Science & Business

Media.
Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77–93.
Raju, D., & Schumacker, R. (2015). Exploring student characteristics of retention that lead to graduation in higher education using data mining models. *Journal of College Student Retention: Research, Theory & Practice*, *16*(4), 563–591.
Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, *24*(4), 567–598.
Seidel, E., & Kutieleh, S. (2017). Using predictive analytics to target and improve first year student attrition. *Australian Journal of Education*, *61*(2), 200–218.
Sothan, S. (2019). The determinants of academic performance: evidence from a cambodian university. *Studies in Higher Education*, *44*(11), 2096–2111.
Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85.
Stratton, L. S., O'Toole, D. M., & Wetzel, J. N. (2008). A multinomial logit model of college stopout and dropout behavior. *Economics of education review*, *27*(3), 319–331.
Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, *45*(1), 89–125.
Vicario, J., Vilanova, R., Bazzarelli, M., Paganoni, A., Spagnolini, U., Torrebruno, A., ... others (2018). Data mining tool for academic data exploitation: selection of most suitable algorithms. *ERASMUS+ KA2/KA203*.
Von Hippel, P. T., & Hofflinger, A. (2020). The data revolution comes to higher education: Identifying students at risk of dropout in chile. *Journal of Higher Education Policy and Management*, 1–22.
Wook, M., Yusof, Z. M., & Nazri, M. Z. A. (2017). Educational data mining acceptance among undergraduate students. *Education and Information Technologies*, *22*(3), 1195–1216.

Appendices

Appendix A: Technical details about data

The dataset: characteristics and variables

PoliMi Information Technology (IT) system collects both dynamic and static data about enrolled students. The former ones are the so-called "digital prints" left in correspondence to some key administrative facts, such as register at exams' sessions, accept or retake grades or pay university's fees. Static data comprises all the information that administrative office registers at the moment of enrolment, such as citisenship, gender or date/place of birth, previous school performance or the university admission test score. The university Administration and IT offices supply the dataset used in the analysis, recording students' information from 2010 to 2019. The number of observations is more than 10 million and each of them represents an administrative event or a student's set of features. The whole dataset is divided into multiple sub-datasets, according to type of information. Hence, data cleaning activity requires to merge the datasets through their linkage with unique encrypted key and to keep only concluded careers, using the student as a unit of analysis, so that we finally consider around $110,000$ students for the analysis. The students' features lastly selected and included into the analysis are summarised in Table 6. The variables are presented and classified in the different groups described in the Section 2.2.

Table 6: Variables' list, description and their belonging to one of the three groups.

group	**Variable's name**	**Description**	**Possible values**
DEM	Gender	Student's gender	1 = male, 0 = female
DEM	Income (range)	Student's contribution fee	Highest income (reference) High income Low income DSU (if the student receives a grant) Unknown income
DEM	Student's origins	Student's Citizenship & Residency	Native Milan: if the student is Italian and live in Milan (reference) Native out Milan: if the student is Italian and live outside Milan Non-Italian abroad: if the student is not Italian and lives outside Italy Non-Italian in Milan: if the student is not Italian, but lives in Milan Non-Italian out of Milan: if the student is not Italian and lives out of Milan
DEM	Access to study age	Student's age at enrolment	From 17 to 50
PRE	PreviousSchool	High school track	Scientific (reference) Classic Technical Other
PRE	Admission Score	Admission test grade	From 60 to 100
ACA	TotalCredits1s/1y	Total credits earned at 1^{st} sem. or 1^{st} year	From 0 to 40 for 1s and from 0 to 80 for 1y
ACA	Attempts 1s or 1y	n. of attempts to pass an exam in the 1^st semester of the 1^st year or in the 1^st year	One: the student attempted the exam once (reference) No: no attempts are done, so the student never attempted the exam More: if the student attempted the exam more than once
ACA	ChangeDegree	Degree program's change	1 if the student changed the degree program within same university, 0 otherwise

A key methodological point consists in the definition of dropout. Indeed, the career of each student is classified as *active*, if the student is actually enrolled, as *suspended*, if the student temporally suspends the career (for example for long trip or pregnancy), as *graduated*, if the student obtained the degree, or *dropout*. This last label identifies students with career's status "definitely closed for a reason different from graduation". In this view, a further clarification is needed: we do not know if dropout students withdraw from Higher Education completely, or just move to another university, changing their academic career. Anyway, this study adopts the university's perspective, providing insights to improve its own retention strategy (i.e. not caring about whether the students succeed in a different institution once moved from it).

The empirical model also includes the different study programs as key variables. Politecnico di Milano offers three kinds of degree courses, grouped in three Schools: Engineering, Architecture and Design. This paper only includes results for the Engineering courses, but the same analysis has been extended for the remaining ones (results available on request from the authors). Specifically, this paper considers the various programs within the School of Engineering as an important element, to detect whether dropout is systematically different across programs. The model considers this important feature of the dataset: its hierarchical structure (students nested into programs) — see details in Section 3.2.

Some evidence from descriptive statistics

This paragraph provides an overview of main information coming from descriptive statistics of variables (See Table 7). In general, students at PoliMi are mainly male (77.7 per cent), Italian (95.1 per cent) and holding a scientific degree from secondary schools (72.4 per cent). The academic path is, in the 17.4 per cent of the cases, not linear, registering a change in their degree program within the university - representing a quite high internal mobility. It is relevant to note that the first year is quite similar to all the courses, allowing students to move among them without losing the formative credits already acquired. When looking at the academic careers, the graduated students register 25 credits earned in the first semester of the first year, against the 5 of dropout students. The situation seems more worrying if we look at the first year: in fact, the cumulative amount of credits earned by graduated students is 52, while the dropout ones is 9. To capture the dimensions of dropout phenomenon, the graduated students are 62.6 per cent of the total considered into the analysis; while dropout occupies the remaining 37.4 per cent, which is composed by 21.6 per cent of early dropouts and 15.8 per cent of late ones.

Table 7: Descriptive statistics of variables used in the models.

	Total		**Graduated**		**Dropout**	
	Mean	*St. Dev.*	*Mean*	*St. Dev.*	*Mean*	*St. Dev.*
AccessToStudiesAge	19.215	2.73	18.813	1.708	19.884	3.781
AdmissionScore	72.453	12.526	75.194	10.854	67.619	13.757
Total Credits 1y	36.064	25.986	52.258	15.919	9.093	14.619
Total Credits 1y1s	18.08	13.099	25.549	8.822	5.641	8.947
AvgAttempts 1y	1.441	0.743	1.606	0.499	1.167	0.967
Avg Attempts 1y1s	1.428	0.809	1.575	0.575	1.183	1.048
Gender						
Female	0.223		0.237		0.198	
Male	0.777		0.763		0.802	
Income						
Highest income	0.338		0.274		0.445	
DSU	0.029		0.043		0.005	
High income	0.333		0.401		0.219	
Low income	0.286		0.277		0.301	
Unknown	0.004		0		0.011	
Student's origins						
Native in Milan	0.261		0.262		0.259	
Native out of Milan	0.69		0.712		0.652	
Non-italian abroad	0.019		0.008		0.037	
Non-italian out of Milan	0.013		0.009		0.022	
Non-italian in Milan	0.017		0.009		0.029	
PreviousSchool						
Scientific	0.724		0.782		0.629	
Classic	0.062		0.061		0.065	
Technical	0.158		0.127		0.209	
Other	0.056		0.03		0.098	
ChangeDegree						
FALSE	0.826		0.823		0.83	
TRUE	0.174		0.177		0.17	

Note: Total contains all the students with concluded careers at PoliMi, Graduated contains only graduated students at PoliMi and Dropout contains only dropout students at PoliMi.

Appendix B: Sensitivity analysis: does this early warning system work well?

In Section 3.2, for both early and late dropout prediction, we included in the third group of covariates of the model, i.e. the group regarding the university career, only the information of the first semester of the first year. This choice is driven by the aim of implementing an early warning system able to predict the wright response as soon as possible. When predicting late student dropout, since the students considered in the model are the ones who attended at least three semester of university, it is possible to consider among the covariates the information of the entire first year of career, instead of only the first semester. This choice should improve the predictive performance of the model, at the

cost of identifying the at risk students later in time, i.e. at the end of the first year instead of at the end of the first semester. In the perspective to analyze whether it is worth to "wait", we run multilevel generalised linear models and multilevel random forest, including all the three groups of covariates, but considering *TotalCredits1y* and the *Attempts1y* instead of *TotalCredits1s* and the *Attempts1s*, i.e. the number of credits obtained and the average attempts per exam computed in the first year instead of in the first semester. The response variable is the binary variable that takes value 1 if the student is a late dropout student and 0 if the student graduated. Table 8 reports the comparison in terms of AUC and sensitivity indexes, of the models considering only first semester information and first year information.

Table 8: Area Under the Curve (AUC) and sensitivity index (sens) of Multilevel generalised linear models and Multilevel random forest with the complete set of covariates considering, among the career information, (i) only the first semester and (ii) the first year, for late dropout prediction.

	Set of covariates included in the model	
	demographic info + previous studies academic performance 1semester	demographic info + previous studies + academic performance 1year
Multilevel generalised linear model	AUC: 0.957 sens: 0.824	AUC: 0.980 sens: 0.892
Multilevel random forest	AUC: 0.948 sens: 0.778	AUC: 0.979 sens: 0.890

Results show that, even if the predictive performances increase by adding to the first semester career information the second semester career information, the gain is not so high to justify the waiting until the end of the first year to identify the students at risk. The small difference between the two models predictive performance suggests that the first semester information is by itself very informative and sufficient to quite precisely predict the students dropout probability and, therefore, it is possible to accurately predict the student dropout probability just observing the beginning of the student university career. This result is very important for a practical viewpoint, as it suggests to use the model for proposing supporting interventions to students as soon as at the end of the first semester.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

39/2020 Martinolli, M.; Biasetti, J.; Zonca, S.; Polverelli, L.; Vergara, C.
Extended Finite Element Method for Fluid-Structure Interaction in Wave Membrane Blood Pumps

40/2020 Fresca, S.; Manzoni, A.; Dedè, L.; Quarteroni, A.
Deep learning-based reduced order models in cardiac electrophysiology

38/2020 Sollini, M.; Kirienko, M.; Cavinato, L.; Ricci, F.; Biroli, M.; Ieva, F.; Calderoni, L.; Tabacchi,
Methodological framework for radiomics applications in Hodgkin's lymphoma

34/2020 Antonietti, P.F.; Mazzieri, I.; Nati Poltri, S.
A high-order discontinuous Galerkin method for the poro-elasto-acoustic problem on polygonal and polyhedral grids

35/2020 Morbiducci, U.; Mazzi, V.; Domanin, M.; De Nisco, G.; Vergara, C.; Steinman, D.A.; Gallo, D
Wall shear stress topological skeleton independently predicts long-term restenosis after carotid bifurcation endarterectomy

36/2020 Pellagatti, M.; Masci, C.; Ieva, F.; Paganoni A.M.
Generalized Mixed-Effects Random Forest: a flexible approach to predict university student dropout

37/2020 Fumagalli, A.; Scotti, A.
A mathematical model for thermal single-phase flow and reactive transport in fractured porous media

31/2020 Bernardi, M.S.; Africa, P.C.; de Falco, C.; Formaggia, L.; Menafoglio, A.; Vantini, S.
On the Use of Interfeometric Synthetic Aperture Radar Data for Monitoring and Forecasting Natural Hazards

32/2020 Menafoglio, A.; Sgobba, S.; Lanzano, G.; Pacor, F.
Simulation of seismic ground motion fields via object-oriented spatial statistics: a case study in Northern Italy

33/2020 Centofanti, F.; Fontana, M.; Lepore, A.; Vantini, S.
Smooth LASSO Estimator for the Function-on-Function Linear Regression Model