Early-predicting dropout of university students: an application of innovative multilevel machine learning and statistical techniques

Marta Cannistrà[1], Chiara Masci[2], Francesca Ieva[2], Tommaso Agasisti[1] and Anna Maria Paganoni[2]

1 Politecnico di Milano, School of Management
2 Politecnico di Milano, Department of Mathematics

Abstract. This paper combines a theoretical-based model with a data-driven approach to develop an Early Warning System that detects students who are more likely to dropout. The model uses innovative multilevel statistical and machine learning methods. The paper demonstrates the validity of the approach by applying it to administrative data from a leading Italian university.

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

Acknowledgments. This research stems from an institutional initiative launched by Politecnico di Milano 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. We are grateful to the University's management for their support and encouragement and to the IT Office of Politecnico di Milano for their support in extracting data and pre-processing them. All the eventual errors are our solely responsibility.

> "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, 2018), 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, 2018). 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.

A high incidence of dropout rates in the functioning of the HE system generates equity and efficiency problems. 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 et al. 2008; Brunori et al. 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[1].

Given the problems associated with dropout, a key policy issue is finding ways to understand, predict and prevent this phenomon. A recent trend in this area is the use of Learning Analytics (LA) tools (De Freitas et al. 2015). Advanced techniques, rooted in both the statistical and Machine Learning (ML) domains, can be used 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 used to design 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. The potential consequences of using LA for practitioners are immediate and relevant. Indeed, if the algorithms work well in early predicting dropout, then HEIs' managers can define interventions and courses targeted to specific individuals who are more at risk of leaving the studies without a degree, with the aim of improving their retention.

This paper contributes to this new literature stream and institutional development. We develop innovative methods to formulate predictions of at-risk students early in their academic career, and we test them using administrative data from Politecnico di Milano (PoliMi), Italy. 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 to the student journey (her initial administrative record, exams, etc.).

This paper answers the following research question: How do alternative algorithms' types (ML

[1] An important note is needed here. Dropout represents a net waste of resources in the cases in which students leave university, but sometimes they do so for switching major or university. In this latter case, the effect is not a net waste of resources for the society, but only for the abandoned university. The argument holds its validity then, although its application is dependent upon the specific definition of dropout. In this paper, we consider the viewpoint of the single university involved (see the section about Methodology and data).

vs Generalised Linear Models) perform in predicting actual dropout and how do we interpret their results? This paper considers answering this research question a condition paving the way for subsequent interventions to be realized in supporting students who are at-risk of dropping out university.

This study innovates the current state-of-the-art of the field in two main directions. First, we develop a comprehensive approach for studying dropout in a data analysis perspective, complementing the application of techniques to the existing data with a conceptual framework 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 (Fun Li et al. 2013; Seidel and Kutieleh, 2017; Vicario et al. 2018; Korhonen and Rautopuro, 2019; Sothan, 2019; Barbu et al. 2019). 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 (De Freitas et al. 2015; Leitner et al. 2017). In practical terms, we exploit all the available administrative data about students (demographic, academic performance, prior achievement, a proxy for the socioeconomic status, etc.) for identifying the variables that are mostly correlated to the precision of predicting students' dropout. While we do not select the variables to be used in the algorithms, we use the lenses of a specific conceptual framework about dropout to interpret their validity and conceptual soundness. Second, we compare different algorithms, built following alternative hypotheses and specifications, to test the validity and robustness of a number of statistical and ML methods. In so doing, we rely upon a set of newly developed methods (within the family of mixed models) that take into account the nested structure of data. In particular, the new methods adopted here consider the students within different degree courses, a feature that is decisive if dropout probability depends on the specific course chosen. The results provide evidence about the accuracy and robustness of predictions about the probability that a specific student would actually drop out.

The remainder of the paper is organised as follows. In Section 2, we develop the conceptual framework for deriving the empirical models in the Learning Analytics perspective. Section 3 describes the methods and data. Section 4 reports the main results. Lastly, Section 5 discusses the main implications and general suggestions towards implementing future interventions for helping at-risk students.

2 Academic literature and conceptual framework

2.1. Related Literature

The academic literature distinguishes between two approaches investigating the features of students' dropout: 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 (Spady, 1970; Tinto, 1975; Pascarella and Terenzini, 1980; Cabrera et al. 1990; St John et al. 1996) propose models to show the processes of interactions between students, their characteristics and the institutions that lead to dropout (Tinto, 1975). These approaches consider 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).

An alternative approach deals with data-driven studies, in which students' characteristics are analysed longitudinally to predict dropout or graduation (Kotsiantis et al. 2003; Fun Li et al. 2013; Seidel and Kutieleh, 2017; Vicario et al. 2018; Solís et al 2018; Nagy and Molontay, 2018; Mayra and Mauricio, 2018; Korhonen and Rautopuro, 2019; Sothan, 2019; Barbu et al. 2019; Alban and Mauricio, 2019; Silva et al. 2020; Heredia-Jiménez et al. 2020). The methodological approach to study dropout in HE described in these works is innovative. Indeed, as highlighted by Agrusti et al. 2019, researches on university dropout prediction increased considerably starting from 2017. The applications proposed in literature are various. Starting from the models adopted, ranging from the more traditional logistic regression (Mayra and Mauricio, 2018) to the innovative Machine Learning algorithms (Alban and Mauricio, 2019; Nagy and Molontay, 2018), also the university considered may be one (Heredia-Jimenez et al. 2020) or more (Silva et al. 2020) or with open courses (Kotsiantis, 2003). Moreover, the information considered for predictions may relate to specific students' features, such as only demographics and pre-college information (Heredia-Jiménez et al. 2020; Nagy and Molontay, 2018), or they exploit all possible knowledge about students (Silva et al. 2020). Results show that Machine Learning models often provide accurate predictions, leaving room for further interventions aiming at retaining potential dropout students. Anyway, in the cited cases, researchers are less interested in explaining the phenomenon *per se*, while the focus is on predicting withdrawing students with the highest level of accuracy.

Placing at the mid-way between theory and data driven studies, some research papers show how the Machine Learning approach may be valuable to support the understanding of dropout (Berens et al. 2018; Rodríguez-Muñiz et al. 2019; Del Bonifro et al. 2020; Sandoval-Palis et al. 2020). Del Bonifro et al. (2020) concentrates on the on-time prediction to detect and then help at risk students as early as possible. On practical strand, they consider only the information acquired at the moment of the students' enrolment. Sandoval-Palis et al. (2020) and Rodríguez-Muñiz et al. (2019) enrich the prediction of dropout students with a deep interpretation of the main determinants of withdrawal, with the aim of arriving to the root causes of the problem. Sandoval-Palis et al. (2020) find that students with the highest risk of dropping out are those in vulnerable situations, with low application grades, enrolled in the levelling course for technical degrees. Results of Rodríguez-Muñiz et al. (2019) work show that the influence of personal and contextual variables and the academic performance in the first year represent the main predictors of dropout. Further, this model highlights other interesting factors: the importance of dedication (part or full time), and the vulnerability of the students with respect to their age. Lastly, Berens et al. (2018) supplement traditional administrative data with approximations of learning behavior and student-teacher interactions recalling the Tinto's integration model. Indeed, they adopted registration in online learning platforms, use of the university library, reading behavior data from the online library as well as online activity level.

2.2. Conceptual Framework

The present paper develops a clear conceptual framework for the comprehesion and interpretation of dropout at university. It considers both the educational process and the need of predicting students' outcome as early as possible. In particular, the data-driven approach is substituted with an information-driven modelling, since the 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 et al. 2017).

With the aim of filling the gaps within the two approaches, the conceptual framework proposed here poses its basis on a student's "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 cumulative process where different factors (e.g. investments, environments and genes) intervene. The technology that governs this process is formed by sequential periods influencing each others and resulting in the educational formation of the individual. Contextualizing this framework into our research, we consider educational stages as school cycles: childhood, primary school, middle school, high school (we use “K12” to refer to all school’s grades until the 12th) and university. During each stage, it is possible to gather different types of information about students’ characteristics and performance. 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 income. The key feature of this model is that individual experiences enrich students’ personal timeline. The milestone of the proposed framework relies on the possibility to predict student’s dropout, considering the previous educational stages as input. This conception brings to deal with an optimization problem, facing the trade-off between prediction *accuracy*, which normally improves when adding more features, and the potential *timing to intervene*, that needs to be reduced as much as possible, so with early predictions. This trade-off lays behind the managerial and policy implications of this research: the timing of the prediction is equally important to its accuracy. The incorrect prediction about possible dropouts may lead institutions to promote targeted remedial interventions for wrong students, risking to esclude the real dropouts. On the other side, intuitively, the more information is available, the more accurate is the prediction. Anyway, collecting data on students’ educational path require time, during which students may decide to leave the university. Hence, balancing time of prediction and information collected is an optimization problem for dropout detection. Further, the critical choice is not only related to the time of prediction, but also to the model adopted, which needs to be the one which better optimize the trade off between accuracy and timing.

From an operational standpoint, a *reduced* view of the proposed conceptual framework needs to contextualise it into real-world practice. Our main assumption related to the optimization problem states that the first moment where we are able to predict, with satisfying accuracy, students’ outcome (graduation or dropout) is the end of the first semester of their first year. So, the complete timeline from HE’s 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 (related to first semester of first year).

3 Methodology and data

3.1 The methodological approach for the empirical analysis: overview

When developing a sound methodology for an accurate and timely prediction of student dropout, this paper considers two main methodological challenges and issues.

First, we must take into account that students are nested within different engineering degree courses. This induces a natural source of dependence among students due to the fact that they are enrolled in the same degree course. Since classical regression models assume all observations to be independent and do not take into account any type of latent structure, multilevel regression models (Pinheiro and Bates, 2006; Goldstein, 2011; Agresti, 2018) are adopted. This class of models are suited to handle the hierarchical structure of data, taking into account the induced dependence among observations. Besides modelling this intrinsic data structure, these models disentangle the variability explained by each level of grouping, helping the analyst in understanding the contribution given by each different level to the response.

A second methodological aspect concerns models’ assumptions. Generalised linear models are the most frequently used techniques in the literature to predict student dropout. Nonetheless, they impose a parametric functional form on the association between the covariates and the response that sometimes results to be too restrictive or unrealistic for describing complex data. For this reason, we compare the results of generalised linear models with the ones obtained applying ML techniques, such as Classification and Regression Trees (CARTs) and Random Forest (RF) (Hastie et al. 2009; Breiman, 2001). These are flexible methods able to investigate non linear associations among the covariates and the response and to model interactions among them. Recent developments in this context allow classification trees to handle hierarchical data: in Fontana et al. (2021), the authors propose a method to fit generalised mixed-effects regression trees (GMET), while, in Pellagatti et al. (2021), the authors develop a new method to fit generalised mixed-effects random forest (GMERF). These methods have the strength and the flexibility of ML techniques, still maintaing the ability to model the nested structure of data. Moreover, although the literature already investigates the main determinants of student dropout (Fun Li et al. 2013; Seidel and Kutieleh, 2017; Vicario et al. 2018; Korhonen and Rautopuro, 2019; Sothan, 2019; Barbu et al. 2019), their estimated effects might vary across methods (i.e., parametric and nonparametric methods). Linear models provide a coefficient for each covariate, that measures the increase in the response for one unit increase in the covariate. Tree-based methods provide a different type of result that consists in the quantification of each covariate's *importance* (measured adopting different criteria) and in the *estimation of the functional form* that marginally links each covariate to the response. In this perspective, we are interested in comparing the predicive power and the interpretative potential of the aforementioned types of methods, considering these two methodological reflections in the analyses of results.

3.2 The methodological approach: mathematical details

We recall now the basics of multilevel models, specifying their modelling both for generalised linear models and 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 course, 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 course and $N = 20$ is the total number of engineering degree courses at PoliMi. 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:

$$\begin{aligned} &\mu_{ij} = \mathbb{E}[Y_{ij}] \qquad\qquad j = 1, \dots, n_i, \quad i = 1, \dots, N \\ &g(\mu_{ij}) = \eta_{ij} \\ &\eta_{ij} = \textstyle\sum_{k=1}^{K+1} \beta_k x_{ijk} \end{aligned} \tag{1}$$

where $\mu_{ij} = p_{ij}$. p_{ij} is the probability that student j within degree course i drops, $g(\mu_{ij})$ is the logit link function, i.e. $g(\mu_{ij}) = logit(\mu_{ij}) = logit(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right)$. K is the total number of predictors, $\boldsymbol{\beta}$ is the $(K+1)-$dimensional vector of coefficients and $\mathbf{x}_{ij}$ is the $(K+1)-$dimensional vector of the covariates (including 1 for the intercept) relative to the (ij)-th observation. This modelling assumes that all 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 courses), the Generalised (logistic) Linear Multilevel Model, GLMM (Agresti, 2018), considering two levels, takes the following form:

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

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 coefficients. 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 course. To verify whether the hierarchical structure taken into account by multilevel models improves dropout predictions, we compare multilevel models' performances with the ones of models not considering degree courses and of models including the degree courses information as a categorical student-level covariate (see Tables A1 and A2 in Annex).

Moving now to a ML setting, the GMET modelling (Fontana et al. 2021) basically substitutes the linear fixed-effects part in Eq. (2) with a tree structure:

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

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 boxes are automatically built by tree in order to minimize the variability within them and maximize the variabilty between them. The absence of a specific functional form makes this method very flexible and able to better model interactions among the covariates. GMET, as standard CARTs, makes an intrinsic selection of the covariates: not all covariates are used in the splits that define the tree, but only the ones that result to be relevant. The covariate used in the first split is the most relevant one and so on. Moreover, different branches of the tree can be defined by different subsets of covariates and this building process reveals the interaction structure among covariates[2].

Similarly, GMERF (Pellagatti et al. 2020) substitues the standard tree $f(\mathbf{x}_{ij})$ in Eq. (3) with a RF, that is an ensemble of trees. RF 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 significant and

[2] It is worth to recall that relevance of covariates and threshold values in the splits are automatically identified by the tree, standing on certain input parameters.

correlated ones (Hastie et al. 2009). Therefore, the advantage of RFs is twofold: they reduce the model variance and they handle the presence of highly correlated covariates, disentangling their associations with the response variable. RFs provide the importance ranking of the covariates in predicting the response, measured as the mean decrease in Gini index – obtained by adding up the total amount that the Gini index is decreased by splits over a given predictor, averaged over all trees of the ensemble (Raileanu and Stoffel, 2004). Moreover, related partial plots displays the marginal association, estimated by GMERF, between each covariate and the response, averaging out the effect of all other covariates.

In the light of these methodological aspects, we expect tree-based methods to identify a similar set of significant covariates. Nonetheless, our main interest is not this one, but it regards two other aspects. The former is the quantification and the qualification of the estimated associations between relevant covariates and the response, compared across different methods. In particuar, we compare results interpretability and releasability. The latter is the quantification of the effect that different assumptions on fixed effects have on the models predictive power.

In this light, standing on the proposed methods and on the different usages we propose about the degree courses information, we run 6 different empirical methods, listed in Table 1.

[Table 1 near here]

3.3 Application – data about Politecnico di Milano

Politecnico di Milano (PoliMi) is the best-ranked Italian public university, and trains students in Engineering, Architecture and Design majors. PoliMi counts around 46,000 students in 2019/2020 in Bachelor and Master courses, among which almost 35,000 in Engineering. This study investigates the phenomenon of student dropout at PoliMi (with specific reference to Engineering bachelor students) and develops a method to early predict it, making a further distinction between *early* and *late* dropout. To clarify the application, a dropout definition is needed: dropout occurs when the student leaves PoliMi for a reason different from graduation. In particular, early dropout occurs when the student drops within the 3rd semester after enrolment[3], while late dropout occurs when the student drops later on. Taking the insitutional standpoint, it is not specified whether the student drops from educational system in general, or he/she shifts major. As stated by Tinto (1982, 2015), it is a matter of perspective and different interests between students, who aim at obtaining a degree, and institutions, which aim at retaining their students. The choice of distinguishing between early and late dropout is motivated by our interest in investigating the determinats of these two types of dropout, that might be potentially different. We expect drivers of an early dropout to be different from the ones of a late dropout. Therefore, each classification model will consider as outcomes of interest early dropout *versus* graduate and late dropout *versus* graduate.

The Information Technology (IT) system of the university 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

[3] We chose this threshold because the third semester after the enrolment represents the deadline for students to enrol in the second academic year.

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. The students' features lastly selected and included into the analysis are summarised in Table 2, divided into demographic, previous studies and academic information.

[Table 2 near here]

Our final sample includes all concluded careers (for dropout or graduation) of students enrolled in an engineering degree course between a.y. 2010/2011 and a.y. 2015/2016. This sample counts 31,071 concluded careers of students, 62.7% of which are graduated, 21.7% are early dropout and 15.6% are late dropout. For both early and late dropout prediction, we train our models on a training set, that is composed by randomly selected 70% of the sample, while the test set is composed by the remaining 30%. In particular, in our models, $Y_{ij} = 1$ when student j within degree course i drops, 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 contains all student-level characteristics shown in Table 2. When we take into account the degree courses information as a categorical student-level variable (*Models 1b*, *2b* and *3b* of Table 1), 19 dummy variables are included. Each dummy variable represents the belonging to one degree course with respect to the reference one (the first one in alphabetic order). When running multilevel models, i.e. when we take into account the hierarchical structure of students nested within degree courses, we include in the random effects part a random intercept, i.e.

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

where b_i is the value-added given by the i −th degree course to the dropout probability (either early or late, depending on the model setting): if b_i is negative, students within the i −th degree course are on average less likely to drop with respect to the others; while, if b_i is positive, students within the i −th degree course 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 (*Model 1c*), to a classification tree in the case of a multilevel classification tree (*Model 2c*) and to a random forest in the case of a multilevel random forest (*Model 3c*). In order to compare the performance of the fitted models, we compute two types of indexes: (i) the Area Under the ROC Curve (AUC), that provides an aggregate measure of performance across all possible classification thresholds; (ii) accuracy, sensitivity ans specificity indexes. Among the set of these performance indexes, we are mainly interested in the sensitivity, because we aim at finding the model that better identifies the at-risk students, i.e., the model with highest sensitivity.

4 Results

We run the 9 models presented in Table 1, for both early dropout *versus* graduated and late dropout *versus* graduated[4]. We analyze the results from two perspectives, recalling the methodological aspects presented in Section 3.1. First, we compare the models' performance, highlighting the main differences between hierarchical and non-hierarchical models and between statistical and ML ones. Then, we compare the types of information about the dropout phenomenon extracted from the proposed models in order to deepen the related mechanisms.

4.1 The performance of the empirical models – overview

The first set of results from the empirical analyses are reported in Tables 3 and 4, which cointain the predictive performances, measured in terms of AUC, sensitivity, accuracy and speificity, of the fitted models, for early and late dropout prediction, respectively. All models' predictive performances are very high, both for *early* and *late* dropout. The lowest value of AUC is 0.8714 and it is reached by the simple tree for predicting late dropout, while the highest one is 0.9615 and it is reached by the GLMM for predicting early dropout. All other models' AUC range between these two values. In particular, classification trees have always slightly lower predictive power than GLM and RF, that, instead, have very similar performances. This difference is more pronounced for early than for late dropout and decreases when considering multilevel models. For both linear and tree-based models, taking into account the degree courses students are enrolled in (both as a dummy variable and by employing a multilevel model) increases their performances, with multilevel models having the highest peak (see Tables 3 and 4). Strengthened by this evidence, we retain multilevel models to be extremely informative in this application. Besides providing the best performance, they fit the real nested structure of students and, especially, they provide interpretable information about the heterogeneities across degree courses (see Section 4.2).

[Tables 3 and 4 near here]

4.2 Understanding and interpreting students' dropout – findings from multilevel generalised linear model and tree based methods

We now focus on the interpretation of the results, reflecting on the various types of information gathered from the proposed models output, adopting a student-level and course-level perspectives. We are interested in investigating whether these methods, that lead to slightly difference performance, give supplemental insights about the dropout phenomenon. In the light of the results shown in Section 4.1, we focus on the multilevel models output, that we retain to be the most informative.

4.3 Individual level factors associated with dropout

Table 5 reports the results of the GLMM (*Model 1c*), both for early and late dropout, respectively[5]. Not significant covariates are removed from the final model using a step-by-step procedure.

[Table 5 near here]

[4] In the early dropout analyses, late dropout students are excluded from the sample and vice-versa.

[5] Tables in Annexes A1 and A2 report detailed results of *Models 1a*, *1b* and *1c*, for early and late dropout, respectively. The association between student-level covariates and the response remains coherent across the models.

Several interesting observations emerge from the associations between student-level information and dropout probability. First, some differences are related to personal characteristics and background. Males are more likely to late drop out than females; although the literature did not reach an agreement about the direction of gender differences in HE dropout, some studies found that female students drop less than their male counterparts (Johnes & McNabb, 2004). Native Italians off-site (i.e. not living in Milan) are more likely to early drop than Italians in-site, perhaps suggesting that commuters and/or students who moved for studying reasons could have encountered additional obstacles to regular academic activities. Non-Italian students are more likely to late drop than native Italians in-site, all else equal – this finding echoes a similar one reported by Meggiolaro et al. (2017) for another Italian university. Students starting their careers at PoliMi at an older age than the average, are more likely to late drop, potentially indicating that these students have a non-linear educational trajectory until their starting moment at PoliMi (for example, they could be students who repeated a grade during secondary education, so are intrinsically more at-risk). Students who attended Other and, especially, technical high schools are more likely to late drop than the ones who attended academic, scientific high schools. This evidence corroborates the heterogeneity across students with different educational background; in Italy, students can attend academic, technical or vocational secondary education, a practice that can hinder equality of opportunities, including later academic success (Brunello and Checchi, 2007).

Differences in dropout probability are also associated with students' previous academic performances. The higher is the admission test score at PoliMi, the higher is the probability of students early dropout, but the lower is the probability of students late dropout. While the negative association with late dropout confirms the right "signalling" effect of the entry test, the positive assocaion with early dropout is quite anomalous. There could be several reasons for this result. It can be the case that high admission scores encourage less motivated students to enrol at PoliMi, then it could happen that they get recruited into the labor market early on or they change university. Further, the higher is the number of credits obtained at the first semester, the lower are both the early and late dropout probabilities, suggesting that students with a good (regular) early start benefit of less risks later on. Students doing more than one attempt 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. These are students who try to pass exams with strong commitment (so they do not drop immeditely), but then are more likely to drop out later if their performance continues struggling. 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; these are students who almost immediately find strong difficulties, and do not even show up at first exams, becoming unable to fill the gap later in their career[6]. Lastly, students with more disadvanatged background (as measured through the income group) are less likely to early drop than their more advantaged counterparts. Also, DSU students (e.g. with a study grant) are less likely to late drop, meaning that socioeconomic background still plays a role in dropout (Rodriguez-Hernandez et al. 2020). It is worth to notice that this finding is conflicting to the one identified by the majority of the worldwide literature, that sees students from disadvantaged backgrounds facing a higher risk of dropping out, but is in line with previous findings on the Italian case (Belloc et al. 2010). A possible explanation could be that, with respect to the majority of students that are in the highest income range and have a wide range of opportunities,

[6] We are aware that there could be a portion of students who do not take any attempts because they have already decided to drop, creating a potential endogeneity issue in studying the phenomenon. In order to check the robustness of our results and to avoid this potential confounding factor, we re-run our linear models for predicting early dropout excluding from the sample those students who did not take any attempts at the first semester. Results, reported in Table A3, confirm that student characteristics associated to the dropout probability, together with models predictive performance, remain quite unchanged (AUC indexes are slightly lower when excluding zero attempts students).

more disadvantaged students are more motivated or feel financial pressures. Their choice to enroll at university may request sacrifices to their family, spurring them to commit. Moreover, those disadvantaged students who decide to enroll at PoliMi are somehow already self-selected and more motivated than average.

Regarding tree-based methods, Figure 1 reports the fixed-effects trees estimated by GMET, for both early and late dropout, respectively. The number of credits the student obtains at the first semester results to be the most important variable to predict both early and late dropout probability. In particular, this is the only covariate used to build the trees. This result helps us in further understanding the dropout phenomenon. GMET output reveals that, by using the number of total credits as single fixed-effects covariate, we build a classificator that performs very close to much more complex models. This evidence is also corroborated by the variables importance ranking shown in Figure 2, obtained by GMERF. Variables importance rankings in Figure 2 confirm that, for both early and late dropout prediction, the number of total credits obtained at the first semester is the most important covariate, and also, it way distance itself, in terms of importance, from the other covariates. The second covariate of the ranking adds very low information to the prediction with respect to it, and so on so forth. The only other covariate that significantly affects the estimates of early dropout probability is the number of attempts.

Besides this clear and interpretable result regarding the covariates' importance, Figure 3 reports the partial dependence plot of the most important covariate selected by GMERF, i.e. the number of credits. Partial dependence plots show the association between the selected covariate and the response, estimated by GMERF net to the effect of all other covariates. In the perspective of investigating the type of association between the single covariate and the response, this graphical tool is extremely informative since it shows the functional form that links the covariate to the response, estimating it directly from the data without imposing any parametric assumption on it. Panels (3a) and (3b) of Figure 3 show that the associations between the number of credits obtained at the first semester and both early and late dropout probability are approximately linear. Being the number of credits obtained at the first semester the most important variable and having it a linear association with the response, it is reasonable to observe similar performances in GMERF and GLMM.

Although the early academic performance results to be the most significant determinant of student dropout probability, some differences in the dropout probability still emerges between students with different origin, gender or previous study. This finding suggests that there are other unobservable factors at play, connected to student origins and previous studies, that lead to significant differences in the dropout probability.

4.4 Dropout differences across degree courses

Besides the information about student-level characteristics, multilevel models give easily interpretable insights about the nested structure, i.e. the degree courses effect. Standing on the predictive performance of the models and on the coefficients significance, both early and late dropout probabilities vary across engineering degree courses. The Variance Partition Coefficient (VPC) is a common index computed in the multilevel model framework to quantify the portion of variability in the response explained at the highest level of grouping. In our case study, VPC quantifies the portion of variability in student dropout that is explained at the degree courses level. Regarding early dropout, VPCs measure 0.1063 for GLMM (*Model 1c*), 0.0857 for GMET (*Model 2c*) and 0.1803 for GMERF (*Model 3c*). For late dropout, VPCs measure 0.0852 for GLMM (*Model 1c*), 0.1193 for GMET (*Model 2c*) and 0.1276 for GMERF (*Model 3c*). These percentages are not negligible, suggesting that there are significant differences in dropout dynamics across degree courses. Random intercepts estimated by multilevel models represent the value-added (positive or negative) of the 20 degree courses to the dropout probability of their students[7]. These estimates are graphically reported in Figure A4 in Annex.

Differences among degree courses might be due to various aspects, as for example heterogenous quality and difficulty and/or structural differences or movement of students across courses. Available data do not allow investigating these mechanisms more in details, and this topic deserves further attention in the future.

4.5 The number of credits obtained at the first semester: the real milestone

Results of the empirical models confirm that the most powerful predictor of both early and late student dropout is the number of credits the student obtains at the first semester of the first year of career. The initial student performance at the university results to be decisive for the career and observing the number of credits obtained by each student at the first semester gives by itself a very good indicator of the student dropout probability. This does not mean that other characteristics are not relevant, especially, considering that all students characteristics antecedent to the enrolment are somehow partially endogenous with the number of credits obtained at the first semester. Type of previous studies, nationality, residence, admission score and income are potential predictors of the student early academic performance.

In order to investigate this association, we regress the number of credits obtained at the first semester against student characteristics antecendent to the enrollment. We consider all students in the sample, i.e. enrolled at PoliMi between 2010 and 2015. In particular, we dichotomise the variable *TotalCredits1s* in a binary variable called *Credits_01* that takes value 1 if *TotalCredits1s* < 7.5 and value 0 if *TotalCredits1s* ≥ 7.5. We chose the threshold value 7.5 identified by the GMET model (*Model 2c*) as the most important split to differentiate graduate *vs* early drop students.

Results of the generalized linear model are reported in Table 6 (the model's AUC values 0.733). All student characteristics antecedent to the enrolment result to be significant for predicting the amount of credits (low or high) obtained by the student at the first semester. This result somehow confirms that students' features are intrisecally and structurally dependent each other, confirming the conceptual framework exposed in Section 2.

[Table 6] near here

5 Discussion, implications and concluding remarks

The results presented in Section 4 can be summarized and commented by answering the research question of this paper.

First, we find a number of factors and variables that are associated with likelihood of dropout, classifiable in the two broad categories of (i) personal background and (ii) previous and early

[7] The technical and mathematical details about the computation of degree courses' effects are reported in Pinheiro and Bates (2006) and Pellagatti et al (2020).

academic performance. Broadly speaking, information belonging to the latter group is more statistically relevant for predicting dropout. All the models tested in this paper performs very well in identifying students who are at risk of dropout, and this is especially true for methods based on multilevel modelling. The major validity of multilevel approaches suggests how sorting across different majors, as well as structural differences across the majors themselves, is an important factor affecting the decision of students to drop out.

Second, the use of tree-based classification methods highlights how the formative credits obtained in the first semester is by far the most important factor associated with dropout. The visualization by means of partial plots identifies the relationship between credits and dropout as an almost linear one. This linear correlation explains why machine learning models (i.e., random forests) do not outperform multilevel ones in correctly predicting students who will droput. Indeed, ML techniques 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 the case presented here, the situation is partly different.

Third, the importance associated with early performance to influence dropouts calls for a renovated attention to explore the determinants of first-semester performance. In the paper, we present some exploratory analyses that are able to explain a significant portion of variability across students in this early performance. This type of analysis can help in identifying the at-risk students even before they obtain their first academic results.

The findings hold a number of policy and managerial implications. The ability to early predict students’ dropout is crucial for targeting interventions in a very personalized way. For example, once identified the at-risk students with sufficient precision, it is possible to develop individualized tutoring systems – eventually, with the support of technology. This is a promising direction to develop a fruitful integration between institutional research and student support systems. In this vein, the results presented in the paper are encouraging. Indeed, they provide sufficient evidence about the positive performance of the statistical and machine learning methods employed for the prediction of students’ results. Such predictions happen soon enough in time to develop remedial interventions, which can sustain the difficulties of students in early stages of their higher education path.

Lastly, a discussion about the limitations of this work and future research is needed. The ML methodologies that we chose to address our research question can handle a binary response, but not a multi-category one. To the best of our knowledge, while linear mixed-effects models have been developed also for multinomial responses (Hadfield 2010). This is not the case for mixed-effects trees and RF, that, when dealing with hierarchical observations, can handle only continuous or binary responses. Because of this limitation, we implemented two different models to estimate early and late dropout probability instead of considering a unique multinomial mixed-effects models with a three categories response (i.e., early dropout, late dropout and graduate). The multinomial approach would be of interest since it would allow to include the entire set of observations, i.e. students, in the multinomial model instead of excluding late dropout students when studying early dropout phenomenon and vice-versa, introducing a natural bias in the model. In this perspective, future research will be devoted to the identification of flexible models, able to handle hierarchical observations and multinomial responses.

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.

Alban, M., & Mauricio, D. (2019). Neural networks to predict dropout at the universities. *International Journal of Machine Learning and Computing*, *9*(2), 149-153.

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). Retrieved from https://www.anvur.it/rapporto-biennale/rapporto-biennale-2018.

Agrusti, F., Bonavolontà, G., & Mezzini, M. (2019). University Dropout Prediction through Educational Data Mining Techniques: A Systematic Review. *Journal of E-Learning and Knowledge Society*, *15*(3), 161-182.

Azcona, D., I-Han Hsiao, & Alan F Smeaton. (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., Kawala-Janik, A., Prada, M., Dominguez, M., Spagnolini, A., et al. (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.

Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. Schumpeter Discussion Papers, No. 2018-006, University of Wuppertal, Schumpeter School of Business and Economics, Wuppertal, http://nbn-resolving.de/urn:nbn:de:hbz:468-20180719-085420-5

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,.

Brunello G. & Checchi, D. (2007). Does school tracking affect equality of opportunity? New international evidence. *Economic Policy*, 22(52), 782-861.

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

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.

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

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020, July). Student dropout prediction. In *International Conference on Artificial Intelligence in Education* (pp. 129-140). Springer, Cham.

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. (2021). Performing Learning Analytics via Generalised Mixed-Effects Trees. *Data*, *6*(7), 74.

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*, volume 922. John Wiley & Sons.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of statistical software,* 33(1): 1-22.

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

Heredia-Jiménez, V., Jiménez, A., Ortiz-Rojas, M., Marín, J. I., Moreno-Marcos, P. M., Muñoz-Merino, P. J., & Kloos, C. D. (2020). An early warning dropout model in higher education degree programs: A case study in Ecuador. In *LAUR@ EC-TEL* (pp. 58-67).

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

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2003, September). 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). Springer, Berlin, Heidelberg.

Johnes, G., & McNabb, R. (2004). Never give up on the good times: student attrition in the UK. *Oxford Bulletin of Economics and Statistics*, *66*(1), 23-47.

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

Kin Fun Li, Rusk, D., & Song, F. (2013). Predicting student academic performance. In *2013 Seventh International Conference on Complex, Intelligent, and Software Intensive Systems*, pages 27–33. IEEE.

Mayra, A., & Mauricio, D. (2018, April). Factors to predict dropout at the universities: A case of study in Ecuador. In *2018 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1238-1242). IEEE.

Meggiolaro, S., Giraldo, A. & Clerici, R. (2017). A multilevel competing risks model for analysis of university students’ careers in Italy. *Studies in Higher Education*, 42(7), 1259-1274.

Nagy, M., & Molontay, R. (2018, June). Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd international conference on intelligent engineering systems (INES)* (pp. 000389-000394). IEEE.

OECD. Education at a glance 2019: OECD indicators. 2019. 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. *Statistical Analysis and Data Mining: The ASA Data Science*, 14(3), 241-257.

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.

Rodriguez-Hernandez, C. F., Cascallar, E., & Kyndt, E. (2020). Socio-economic status and academic performance in higher education: A systematic review. *Educational Research Review*, *29*, 100305.

Rodríguez-Muñiz, L. J., Bernardo, A. B., Esteban, M., & Díaz, I. (2019). Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques?. *Plos one*, *14*(6), e0218796.

Sandoval-Palis, I., Naranjo, D., Vidal, J., & Gilar-Corbi, R. (2020). Early Dropout Prediction Model: A Case Study of University Leveling Course Students. *Sustainability*, *12*(22), 9314.

Seidel, E. & Kutieleh, S. (2017). Using predictive analytics to target and improve first year student attrition. *Australian Journal of Education*, 61(2):200–218.

Silva, J., Matos, L. F. A., Mosquera, C. M., Mercado, C. V., González, R. B., Llinás, N. O., & Lezama, O. B. P. (2020). Prediction of academic dropout in university students using data mining: Engineering case. In *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies* (pp. 495-500). Springer, Singapore.

Solís, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018, July). Perspectives to predict dropout in university students with machine learning. In *2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI)* (pp. 1-6). IEEE.

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.

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

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125.

Tinto, V. (1982). Defining dropout: A matter of perspective. *New Directions for Institutional Research*, *1982*(36), 3-15.

Tinto, V. (2017). Through the eyes of students. *Journal of College Student Retention: Research, Theory & Practice, 19*(3), 254-269.

Vicario, J., Vilanova, R., Bazzarelli, M., Paganoni, A. M., Spagnolini, U., Torrebruno, A., Prada, M., Morán, A., Dominguez, M., Pereira, M. J., et al. (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*, pages 1–22.

Wook, M., Yusof, Z. M., & Ahmad Nazri, M. Z. (2017). Educational data mining acceptance among undergraduate students. *Education and Information Technologies*, 22(3):1195–1216.

Tables and figures

Table 1: Proposed empirical models for analyzing early and late dropout.

Degree courses approach Type of models	Degree courses not considered	Degree courses considered: dummy variable	Degree courses considered: multilevel model
Generalised linear model	Model 1a	Model 1b	Model 1c
Classification tree	Model 2a	Model 2b	Model 2c
Random forest	Model 3a	Model 3b	Model 3c

Note: the table presents the overview of the run models, dividing them according to their typology (linear, tree or random forest) and to the ways of considering the degree courses information (ignored, included as a categorical variable or by a multilevel approach).

Table 2: Student-level variables' list: description and domain

Group	Variable's name	Description	Possible values	Descriptive statistics[8]
Demographic information	*Gender*	Student's gender	*1:* male	76.17%
			0: female	23.83%
	Income (range)	Student's contribution fee	*Highest income* (reference)	33.58%
			High income	32.55%
			Low income	27.97%
			DSU (if the student receives a grant)	3.20%
			DK, Unknown income	2.70%
	Access to study age	Student's age at enrolment	From 17 to 50	19.27 (*IQR: 18.00 – 19.00*)
	Student' origins	Student's Citizenship & Residency	*Native Milan*: if the student is Italian and live in Milan (reference)	*25.54%*
			Native out Milan: if the student is Italian and live outside Milan	*67.57%*
			Non-Italian abroad: if the student is not Italian and lives outside Italy	*3.87%*
			Non-Italian in Milan: if the student is not Italian, but lives in Milan	*1.69%*
			Non-Italian out of Milan: if the student is not Italian and lives out of Milan	*1.33%*
Previous studies and performance information	*Previous Studies*	High school track	*Scientific* (reference)	73.33%
			Classic	6.32%
			Technical	15.97%
			Other	4.38%
	Admission score	Admission test grade	From 60 to 100	72.45 (*IQR: 81.07 – 64.57*)
Academic information	*TotalCredits1s*	Total credits earned at 1st sem. of 1st year	From 0 to 40	17.97 (*IQR: 30.00 – 0.00*)
	Attempts 1s	n. of attempts to pass an exam in the 1st semester of the 1st year	*One*: the student attempted the exam once (reference)	24.17%
			No: no attempts are done, so the student never attempted the exam	*12.94%*
			More: if the student attempted the exam more than once	*62.89%*

Note: The table presents the list of variables adopted in the models with their description and assumed values. When dealing with a categorical variable, we point out the reference level – usually the most populated one.

[8] We provide mean and interquartile range for numerical variables and percentage for categorical variables

Table 3: Area Under the ROC Curve (AUC) and accuracy, sensitivity and specificity indexes of the 9 models run for early dropout *versus* graduated

	Not nested (a)		Dummy (b)		Nested (c)	
Generalised Linear Model (1)	AUC = 0.9576	Acc = 0.9178 Sen = 0.8943 Spec = 0.9234	AUC = 0.9614	Acc = 0.9219 Sen = 0.8913 Spe = 0.9291	AUC = 0.9615 VPC = 0.1063	Acc = 0.9224 Sen = 0.8921 Spec = 0.9296
Classification Tree (2)	AUC = 0.8748	Acc = 0.9342 Sen = 0.7789 Spec = 0.9708	AUC = 0.8748	Acc = 0.9342 Sen = 0.7789 Spe = 0.9708	AUC = 0.9473 VPC= 0.0857	Acc = 0.9118 Sen = 0.9004 Spec = 0.9145
Random Forest (3)	AUC = 0.9512	Acc = 0.9183 Sen = 0.8709 Spec = 0.9294	AUC = 0.9553	Acc = 0.9155 Sen = 0.8898 Spe = 0.9216	AUC = 0.9598 VPC=0.1803	Acc = 0.916 Sen = 0.8966 Spec = 0.9205

Note: The sensitivity is obtained as sensitivity = # true positive / (#true positive + #false negative), 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. ROC curve is a graphical plot that illustrates the diagnostic ability of the classifier system as its discrimination threshold is varied. AUC measures the area under the ROC curve and is equal to the probability that the classifier will rank a randomly chosen dropout student higher than a randomly chosen graduated one (assuming dropout ranks higher than graduate). AUC =1 is the perfect fitting.

Table 4: Area Under the Curve (AUC) and accuracy, sensitivity and specificity indexes of the 9 models run for late dropout *versus* graduated

	Not nested (a)		Dummy (b)		Nested (c)	
Generalised Linear Model (1)	AUC = 0.8977	Acc = 0.8637 Sen = 0.7634 Spec = 0.8855	AUC = 0.9089	Acc = 0.8593 Sen = 0.8003 Spec = 0.8721	AUC = 0.9091 VPC=0.0852	Acc = 0.859 Sen = 0.7979 Spec = 0.8723
Classification Tree (2)	AUC = 0.8714	Acc = 0.8851 Sen = 0.673 Spec = 0.9312	AUC = 0.89019	Acc = 0.8157 Sen = 0.8718 Spec = 0.8035	AUC = 0.9049 VPC = 0.1193	Acc = 0.8393 Sen = 0.8118 Spec = 0.8453
Random Forest (3)	AUC = 0.8897	Acc = 0.864 Sen = 0.7354 Spec = 0.8919	AUC = 0.9016	Acc = 0.8519 Sen = 0.788 Spec = 0.8658	AUC = 0.9065 VPC =0.1276	Acc = 0.8549 Sen = 0.8036 Spec = 0.866

Note: The sensitivity is obtained as sensitivity = # true positive / (#true positive + #false negative), 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. ROC curve is a graphical plot that illustrates the diagnostic ability of the classifier system as its discrimination threshold is varied. AUC measures the area under the ROC curve and is equal to the probability that the classifier will rank a randomly chosen dropout student higher than a randomly chosen graduated one (assuming dropout ranks higher than graduate). AUC =1 is the perfect fitting.

Table 5: Coefficients of GLMMs for early and late dropout prediction

	Dependent variable:	
	Dropout vs. Graduated	
	(Early)	*(Late)*
Gender (ref.: female)		0.616***
		(0.087)
Prev Stud: Classic (ref.: scientific)		-0.186
		(0.131)
Prev Stud: Other (ref.: scientific)		0.266*
		(0.161)
Prev Stud: Technical (ref.: scientific)		0.177**
		(0.08)
Native out of Milan (ref.: Native Milan)	0.341***	0.101
	(0.086)	(0.067)
Non-Italian abroad (ref.: Native Milan)	0.008	0.760**
	(0.457)	(0.36)
Non-Italian in Milan (ref.: Native Milan)	0.209	0.485**
	(0.364)	(0.228)
Non-Italian out of Milan (ref.: Native Milan)	0.002	0.347
	(0.337)	(0.238)
Admission Score	0.015***	-0.006**
	(0.004)	(0.003)
Access to studies age		0.188***
		(0.025)
TotalCredits1.1	-0.228***	-0.171***
	(0.005)	(0.004)
attempts1: more (ref.: one)	-0.682***	0.463***
	(0.096)	(0.089)
attempts1: none (ref.: one)	2.339***	0.973***
	(0.288)	(0.274)
Family Income: DSU (ref.: highest)	-0.332	-0.754***
	(0.286)	(0.264)
Family Income: High (ref.: highest)	-0.222**	-0.1
	(0.094)	(0.077)
Family Income: Low (ref.: highest)	-0.163*	0.116
	(0.097)	(0.078)
Family Income: DK (ref.: highest)	-1.227	-0.773
	(1.031)	(0.514)
Constant	1.103***	-2.665***
	(0.322)	(0.552)
Observations	16,216	15,901
Log Likelihood	2,703.508	-4,019.046
Akaike Inf. Crit.	5,435.017	8,076.093
Bayesian Inf. Crit.	5,542.729	8,221.901

Note: 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.

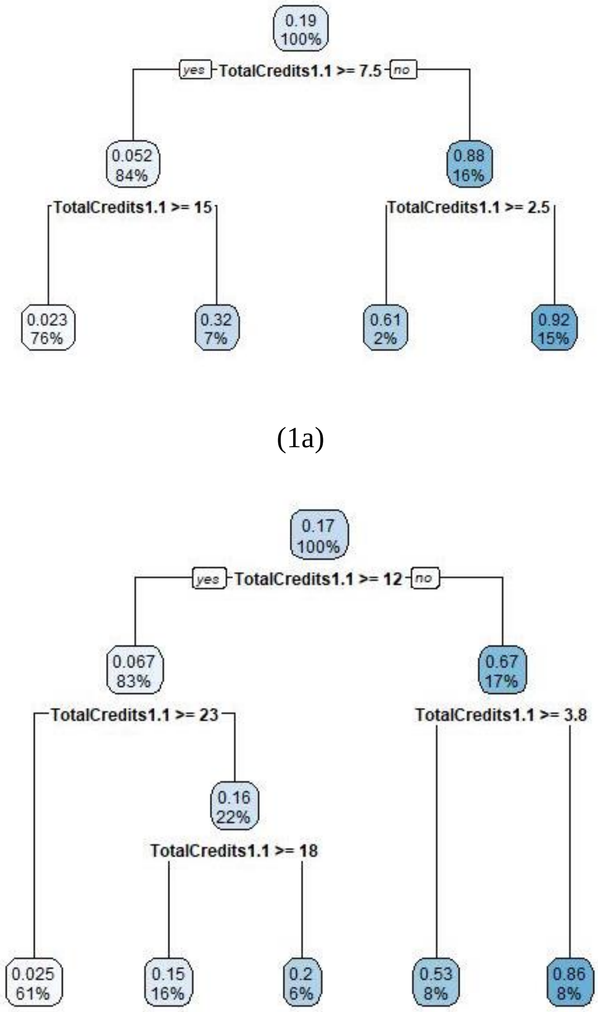
Table 6: Results of the GLM for predicting *Credits_01*, considering all students enrolled between 2010 and 2015

	Dependent variable: Credits_01
Gender (ref.: female)	0.224***
	(0.048)
Previous Studies: Classic (ref.: scientific)	0.061
	(0.077)
Previous Studies: Other (ref.: scientific)	0.527***
	(0.101)
Previous Studies: Technical (ref.: scientific)	0.045
	(0.055)
Native out of Milan (ref.: Native Milan)	-0.011
	(0.043)
Non-Italian abroad (ref.: Native Milan)	0.358
	(0.236)
Non-Italian in Milan (ref.: Native Milan)	0.624***
	(0.164)
Non-Italian out of Milan (ref.: Native Milan)	0.390***
	(0.160)
Admission Score	-0.651***
	(0.023)
Access to studies age	0.316***
	(0.024)
Family Income: DSU (ref.: highest)	-5.809***
	(1.005)
Family Income: High (ref.: highest)	-1.080***
	(0.047)
Family Income: Low (ref.: highest)	-0.743***
	(0.048)
Family Income: DK (ref.: highest)	-14.572
	(97.677)
Constant	-0.993***
	(0.057)
Observations	18,865
Log Likelihood	-8,483.819
Akaike Inf. Crit.	16,997.640

Note: 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*.

Figures

Figure 1: Fixed-effects trees obtained by GMET (*Model 2c*), for early (panel 1a) and late (panel 1b) dropout prediction

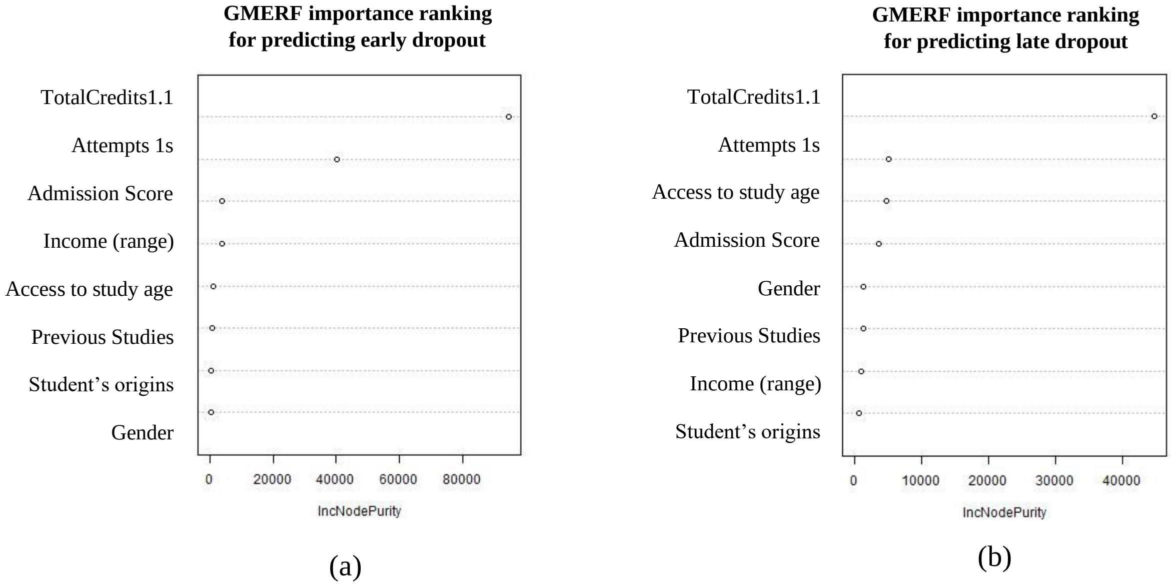


(1a)

(1b)

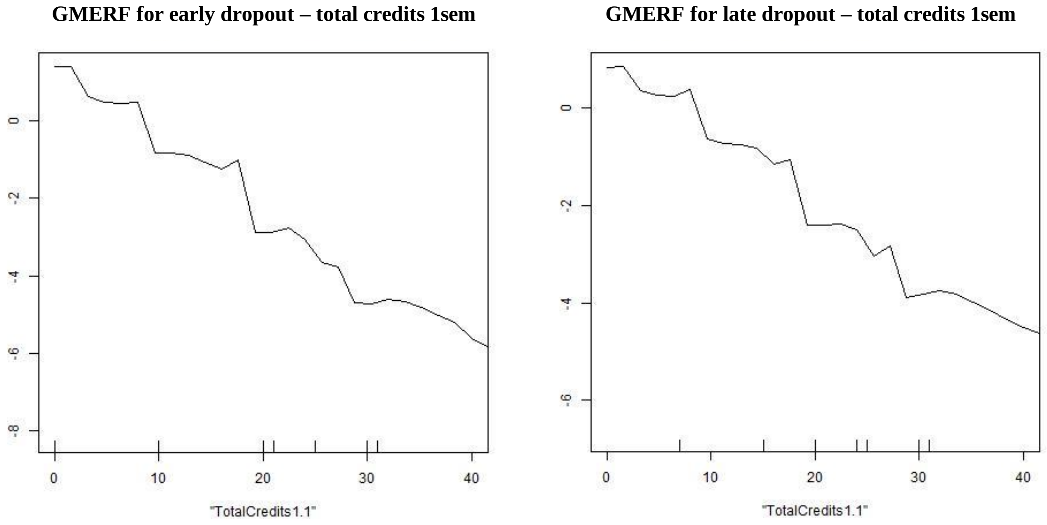
Note: Each node reports the probability of dropout of the percentage of observations (reported below) belonging to the node.

Figure 2: Fixed-effects variable importance plots computed by GMERF (*Model 3c*), for both early and late dropout prediction.



Note: The variable importance measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index.

Figure 3: Variable importance plots of total credits for early (panel 3a) and late dropout (panel 3b), respectively, estimated by GMERF (*Model 3c*).



(3a) (3b)

Note**:** Partial plots report the net effect of the selected covariates on the response (*logit(p)*), after averaging out the effect of all other covariates.

Annexes

Table A1: Complete results of GLMs (*Models 1a, 1b*) and GLMM (*Model 1c*) for the prediction of early dropout vs. graduation.

	Dependent variable:		
	Early dropout vs graduated		
	Logistic models		*Mixed-effects generalized linear model*
	(a)	***(b)***	***(c)***
Native out of Milan	0.338***	0.352***	0.341***
	(0.083)	(0.086)	(0.086)
Non-Italian abroad	0.122	0.004	0.008
	(0.432)	(0.46)	(0.457)
Non-Italian in Milan	0.319	0.202	0.209
	(0.362)	(0.367)	(0.364)
Non-Italian out of Milan	0.094	-0.008	0.002
	(0.33)	(0.338)	(0.337)
Admission Score	0.017***	0.014***	0.015***
	(0.003)	(0.004)	(0.004)
TotalCredits1.1	-0.220***	-0.229***	-0.228***
	(0.004)	(0.005)	(0.005)
attempts1: more	-0.528***	-0.694***	-0.682***
	(0.091)	(0.097)	(0.096)
attempts1: none	2.461***	2.352***	2.339***
	(0.284)	(0.289)	(0.288)
Family Income: DSU	-0.364	-0.325	-0.332
	(0.288)	(0.287)	(0.286)
Family Income: High	-0.257***	-0.217**	-0.222**
	(0.091)	(0.094)	(0.094)
Family Income: Low	-0.172*	-0.158	-0.163*
	(0.094)	(0.097)	(0.097)
Family Income: DK	-1.315	-1.219	-1.227
	(1.038)	(1.034)	(1.031)
Constant	0.802***	1.016***	1.103***
	-0.27	-0.331	-0.322
Control for course enrolment	No	Yes	No
Observations	16,216	16,216	16,216
Log Likelihood	-2,778.236	-2,667.904	-2,703.508
Akaike Inf. Crit.	5,582.472	5,399.808	5,435.017
Bayesian Ing. Crit.			5,542.729

Note*:* 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.*

Table A2: Complete results of GLMs (*Models 1a, 1b)* and GLMM (*Model 1c*) for the prediction of late dropout vs. graduation.

	Dependent variable:					
	Late dropout vs graduated					
	Logistic models				*Mixed-effects generalized linear model*	
	(a)		***(b)***		***(c)***	
Gender Male	0.770***	*(0.083)*	0.607***	*(0.088)*	0.616***	*(0.087)*
Previous Studies: Classic	-0.105	*(0.128)*	-0.188	*(0.132)*	-0.186	*(0.131)*
Previous Studies: Other	0.370**	*(0.152)*	0.265	*(0.161)*	0.266*	*(0.161)*
Previous Studies: Technical	0.408***	*(0.073)*	0.174**	*(0.08)*	0.177**	*(0.08)*
Native out of Milan	0.071	*(0.065)*	0.107	*(0.067)*	0.101	*(0.067)*
Non-Italian abroad	0.950***	*(0.345)*	0.749**	*(0.361)*	0.760**	*(0.36)*
Non-Italian in Milan	0.741***		0.466**	*(0.228)*	0.485**	*(0.228)*
Non-Italian out of Milan	0.606***	*(0.222)*	0.334	*(0.239)*	0.347	*(0.238)*
Admission Score			-0.006**	*(0.003)*	-0.006**	*(0.003)*
Access to studies age	0.206***	*(0.024)*	0.188***	*(0.025)*	0.188***	*(0.025)*
TotalCredits1.1	-0.171***	*(0.003)*	-0.172***	*(0.004)*	-0.171***	*(0.004)*
attempts1: more			0.462***	*(0.09)*	0.463***	*(0.089)*
attempts1: none			0.974***	*(0.274)*	0.973***	*(0.274)*
Family Income: DSU			-0.752***	*(0.264)*	-0.754***	*(0.264)*
Family Income: High			-0.098	*(0.077)*	-0.1	*(0.077)*
Family Income: Low			0.116	*(0.078)*	0.116	
Family Income: DK			-0.779	*(0.514)*	-0.773	
Constant	-3.197***		-2.894***		-2.665***	
	(0.461)		(0.553)		(0.552)	
Control for course enrolment	No		Yes		No	
Observations	15,901		15,901		15,901	
Log Likelihood			-4,169.700		-3,981.816	
Akaike Inf. Crit.	8,361.400		8,037.631		8,076.093	
Bayesian Inf. Crit.					8,221.901	

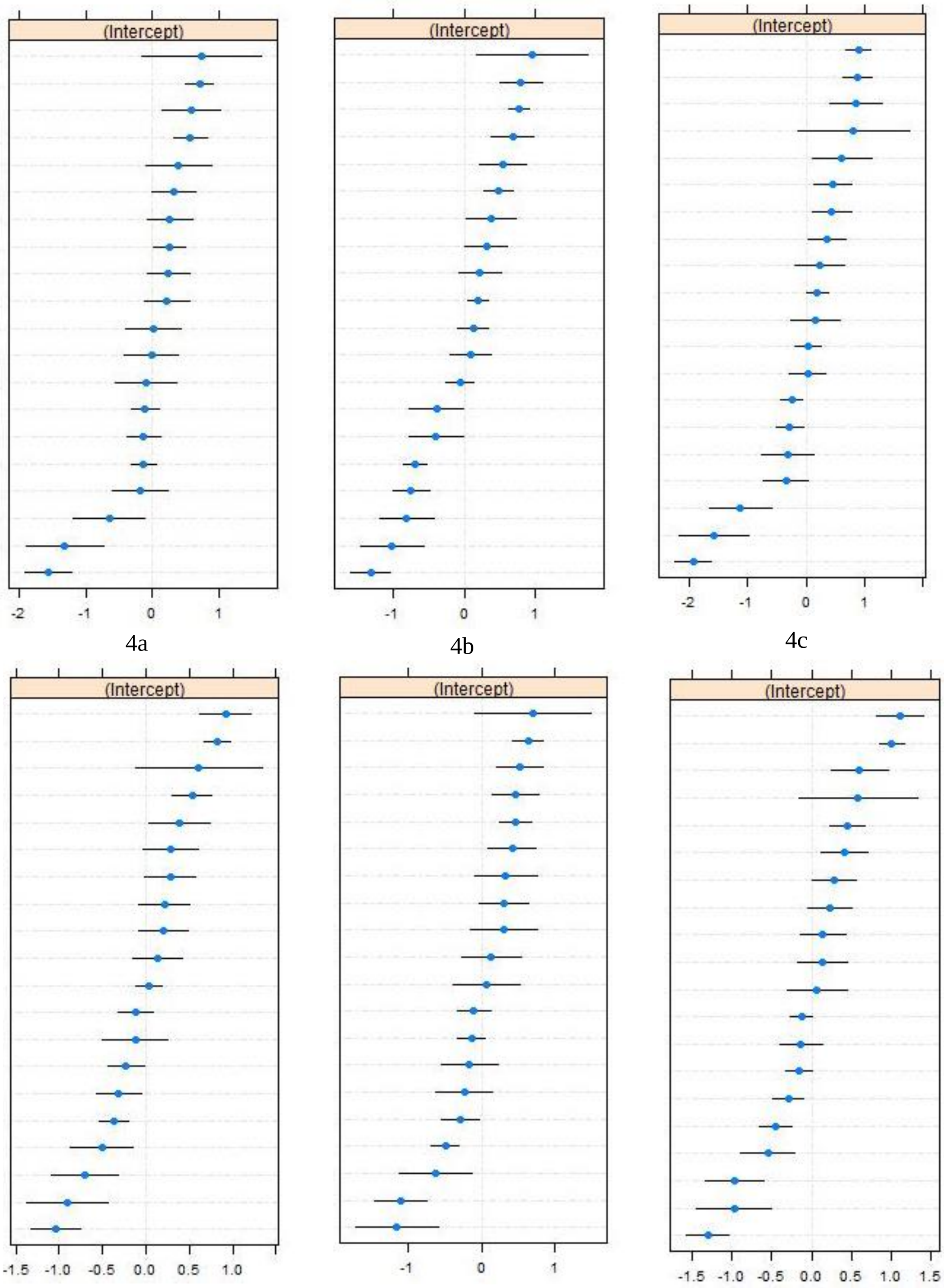
***Note**:* Results are reported in terms of regression coefficients point estimates with their standard deviation (in brackets). Stars represent the statistical significance: **p<0.1; **p<0.05; ***p<0.01.*

Table A3: Complete results from GLM (*Models 1a and 1b*) and GLMM (*Model 1c*) for the prediction of early dropout vs. graduation considering students with more than 0 attempts (i.e. excluding from the analysis those students who did not attempt any exam).

	Dependent variable:		
	Early dropout vs. graduated		
	Logistic models		*Mixed-effects generalized linear model*
	(1)	(2)	(3)
Gender Male	0.191**	0.144	
	(0.09)	(0.097)	
Native out of Milan	0.329***	0.349***	0.329***
	(0.083)	(0.086)	(0.086)
Non-Italian abroad	-0.104	-0.194	-0.187
	(0.492)	(0.535)	(0.526)
Non-Italian in Milan	-0.034	-0.072	-0.118
	(0.391)	(0.389)	(0.38)
Non-Italian out of Milan	0.359	0.24	0.248
	(0.364)	(0.378)	(0.372)
Admission Score	0.011***	0.010**	0.010***
	(0.004)	(0.004)	(0.004)
TotalCredits1.1	-0.215***	-0.225***	-0.226***
	(0.004)	(0.005)	(0.005)
attempts1: more	-0.638***	-0.828***	-0.813***
	(0.09)	(0.097)	(0.096)
Family Income: DSU	-0.531*	-0.488	
	(0.319)	(0.316)	
Family Income: High	-0.221**	-0.187**	
	(0.092)	(0.095)	
Family Income: Low	-0.043	-0.037	
	(0.094)	(0.098)	
Family Income: DK	-1.241	-1.126	
	(1.038)	(1.041)	
Constant	0.997***	1.100***	1.330***
	(0.28)	(0.336)	(0.32)
Control for course enrolment	No	Yes	No
Observations	14,790	14,790	14,790
Log Likelihood	-2,709.975	-2,591.232	-2,632.042
Akaike Inf. Crit.	5,445.949	5,246.463	5,282.084
Bayesian Ing. Crit.			5,350.500

Note: AUC indexes are 0.9258, 0.9311 and 0.9311 for *Models 1a, 1b* and *1c*, respectively. 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.*

Figure A4: Random effects intercepts with relative 95% confidence intervals, estimated by GLMM (*Model 1c),* GMET *(Model 2c)* and GMERF (Model *3c)*. In particular, first line reports the results for early dropout, for GLMM (Panel 4a), GMET (Panel 4b) and GMERF (Panel 4c), respectively. Second line reports the results for late dropout, for GLMM (Panel 4d), GMET (Panel 4e) and GMERF (Panel 4f), respectively.



4a

4b

4c

4a

4b

4c

***Note*:** For anonymity reasons, we do not report degree courses names alongside the estimated rankings. This figure is intended only as a tool to visualize and quantify the variability across degree courses, estimated by the proposed multilevel models.