data

MDPI

Article

Performing Learning Analytics via Generalised Mixed-Effects Trees

Luca Fontana, Chiara Masci *, Francesca Ieva and Anna Maria Paganoni

MOX— Laboratory for Modeling and Scientific Computing, Department of Mathematics, The Polytechnic University of Milan, 20133 Milan, Italy; luca11.fontana@mail.polimi.it (L.F.); francesca.ieva@polimi.it (F.I.); anna.paganoni@polimi.it (A.M.P.)

* Correspondence: chiara.masci@polimi.it

Abstract: Nowadays, the importance of educational data mining and learning analytics in higher education institutions is being recognised. The analysis of university careers and of student dropout prediction is one of the most studied topics in the area of learning analytics. From the perspective of estimating the likelihood of a student dropping out, we propose an innovative statistical method that is a generalisation of mixed-effects trees for a response variable in the exponential family: generalised mixed-effects trees (GMET). We performed a simulation study in order to validate the performance of our proposed method and to compare GMET to classical models. In the case study, we applied GMET to model undergraduate student dropout in different courses at Politecnico di Milano. The model was able to identify discriminating student characteristics and estimate the effect of each degree-based course on the probability of student dropout.

Keywords: mixed-effects models; regression and classification trees; student dropout; academic data; learning analytics

check for updates

Citation: Fontana, L.; Masci, C.; Ieva, F.; Paganoni, A.M. Performing Learning Analytics via Generalised Mixed-Effects Trees. *Data* **2021**, *6*, 74. https://doi.org/10.3390/data6070074

Academic Editor: Kassim S. Mwitondi

Received: 3 May 2021
Accepted: 5 July 2021
Published: 9 July 2021

1. Introduction

The present work is part of the international SPEET project (Student Profile for Enhancing Engineering Tutoring), an ERASMUS$^+$ project aiming to provide a new perspective to university tutoring systems. It intends to extract useful information from academic data provided by its partners[1] and to identify different engineering student profiles across Europe [1]. Our goal was to find out which indicators may discriminate between two different student profiles: *dropout* students, who permanently abandon their Bachelor of Science (BSc) programs, and *graduate* students, who attain the academic qualification. This was motivated by the fact that, across all SPEET partners, almost one student out of two leaves his/her engineering studies before obtaining a BSc degree. If it were possible to know promptly to which profile a student belongs, tutors could improve counselling actions.

Data provided by universities usually include indicators about socio-economic background, and both the current and previous performance data of the students. However, academic success depends on different factors, both internal and external [2]. The dataset we used in our analysis includes information on more than 18,000 BSc students from Politecnico di Milano (PoliMi): it essentially consists of student records, so it does not include all possibly relevant factors. Datasets with similar structures have already been used in recent developments oriented toward performance prediction and detection of future dropouts or students at risk of dropping out [3]. The hypothesis is that background and performance indicators together are enough to identify the students at risk and to draw the attention of tutors, who should complete each student's profile with further information.

In our case study, students were naturally nested within degree-based courses. Further levels of hierarchy are possible, such as programmes within faculties, faculties within universities and finally universities within countries. While investigating the learning process, it is necessary to disentangle the effects given by each level of hierarchy [4]. Indeed, if the clustered aspect of the data is not inspected, it may result in a loss of likely valuable information. Multilevel models take into account the hierarchical nature of data and are able to quantify the portion of variability in the response variable that is attributable to each level of grouping [5]. Generalised linear mixed models (GLMM) fit a multilevel model on a binary response variable, but they impose a linear effect of covariates on a transformation of the response variable [6]. On the contrary, tree-based methods such as the classification and regression tree (CART) model learn the relationships between the response and the predictors by identifying dominant patterns in the training data [7]. In addition, these methods allow a clear graphical representation of the results that is easy to communicate. The goal of our work was to create a novel method, for a non-Gaussian response variable, which is able to preserve the flexibility of the CART model and to extend it to a clustered data structure, where multiple observations can be viewed as being sampled within groups.

This was not the first time that tree-based methods have been adopted to deal with longitudinal and clustered data. In Sela and Simonoff [8], a regression tree method for longitudinal or clustered data was proposed. This method is called the random effects expectation-maximization (RE-EM) tree. Independently, in Hajjem et al. [9] a mixed-effect regression tree (MERT) model was proposed. If clustered observations are considered, these are extensions of a standard regression tree to the case of individuals nested within groups. These methods use observation-level covariates in the splitting process and can deal with the possible random effects associated with those covariates. However, they both deal only with Gaussian response variables, and they are not suitable for classification problems. Our proposed method intends to generalise the RE-EM tree approach, thereby extending its use to different classes of response variables that belong to the exponential family[2]: this should allow one to extend it, for example, to a classification setting. At the same time, this method can deal with the grouped data structure, similarly to traditional multilevel models. As in RE-EM tree estimation, we developed an algorithm that disentangles the estimations of fixed and random effects. That is, an initial tree is built ignoring the grouped data structure, a mixed-effects model is fitted based on the resultant tree structure and a final mixed-effects tree is reported.

Similar methods were proposed in Hajjem et al. [10], Fokkema et al. [11] and Speiser et al. [12], but following different approaches. In Hajjem et al. [10] the MERT approach was extended to non-Gaussian data, and a generalised mixed effects regression tree (GMERT) was proposed. This algorithm is basically the penalised quasi-likelihood (PQL) algorithm used to fit GLMMs, where the weighted linear mixed-effect pseudo-model is replaced by a weighted MERT pseudo-model. In particular, the authors used a first-order Taylor-series expansion to linearise the response variable. In Fokkema et al. [11], the authors proposed the generalised linear mixed-effects model tree (GLMM tree) algorithm, which alternates the estimates of a GLM tree and a mixed-effects model until convergence. Its main distinction from the GMET algorithm is that the GLMM tree algorithm builds on model-based recursive partitioning (MOB, Zeileis et al. [13]), instead of on CART, as GMET does. Lastly, the most recent work was presented in Speiser et al. [12]. The authors developed a decision tree method for modelling clustered and longitudinal binary outcomes. Even if the aim of their model is very similar to ours, their model only handles binary outcomes using a Bayesian GLMM, and it allows a random intercept, but not random slopes. Differently from these cited methods, GMET starts by initialising the random-effects to zero; it estimates the target variable through a GLM (using suitable link functions depending on the response family distribution); builds a regression tree using the estimated target variable as the dependent variable; and then fits a mixed-effects model to estimate the random-effects part, using the fixed-effects part estimated by the tree as an offset.

In the last few decades, learning analytics, and specifically, the topic of dropouts at university, is receiving particular attention. The investigation of the dropout phenomenon within higher education institutions (HEIs) has always been a concern for educators, university managers and policy makers. The academic literature distinguishes between two approaches to investigating the features of this phenomenon: theory-driven and data-driven. The first analyses the reasons and the psychological constructs behind withdrawing decisions, thereby identifying theoretical fundamentals and contributing to a conceptual model to guide the inquiry. Different authors [14–18] proposed models to show the processes of interactions among students, their features and their institutions that lead to dropping out [18]. Basically, their models rely on an interdisciplinary approach to explain the dropout process. In particular, the model considers the interactions between the student and the university environment—individuals are exposed to influences, expectations and demands from a variety of sources (such as courses, faculty members, administrators and peers). The interactions between these two aspects contribute to a student's success or failure in both the academic system and the social system [17]. Hence, these studies focus on the necessity to contextualise the student's educational career in a community structure.

The alternative approach is data-driven. In it, students' characteristics are analysed longitudinally to find the best statistical models predicting dropout or graduation [2,19–21]. In this case, researchers are less interested in explaining the phenomenon per se; the focus is on finding the best performing model in terms of forecasting student withdrawal. The prediction of low performers is increasingly getting the attention of academics, which is attributable to the applicability of remedial learning, which in turn serves the institutional goals of providing high-quality education ecosystems [22]. In addition, the data mining approach to education is quickly becoming an important field of research due to its ability to extract new knowledge from a large amount of student data [23].

The goal behind the present study was the development of a clear theoretical framework, in the midway point between the two approaches, which considers the educational process and the need for predicting students' outcomes as early as possible. We applied the GMET model to the Politecnico di Milano data, collected within the ERASMUS$^+$ SPEET project, thereby identifying which fixed-effect covariates discriminate between dropout and graduate students. Through the GMET model, we relaxed the assumption of linear effects of student-level covariates on their performances, and we identified which interactions relevantly influence dropout status. We included the most common student characteristics in a flexible and interpretable model that takes into account the enrolment in different degree programs. A multilevel model allows one to estimate the degree programme's effect on the predicted probability of obtaining the degree. Machine learning and tree-based methods have been applied in the literature to model student dropout [24–29], but to the best of our knowledge, we are presenting the first time that a *multilevel tree-based method* has been applied to predict student dropout probability.

The paper is organised as follows. In Section 2 we describe the model and methods—the generalised mixed tree algorithm (GMET). In Section 3 we show a simulation study. In Section 4 we describe the PoliMi dataset, we report the application of the proposed algorithm to the case study and we outline the results. Finally, in Section 5 we draw our conclusions.

All the analysis was performed using R software [30]. The R code for the GMET algorithm and for all the simulations is available in Supplementary Materials Data S1.

2. Model and Methods

In this section, we present the proposed generalised mixed-effects tree model (Section 2.1) and the algorithm for the estimation of its parameters (Section 2.2).

2.1. The Generalised Mixed-Effects Tree Model

We start by considering a generic GLMM. This model is an extension of a generalised linear model that includes both fixed and random effects in the linear predictor [6]. Therefore, GLMMs handle a wide range of response distributions and a wide range of scenarios where observations are clustered into groups rather than being completely independent. For a GLMM with a two-level hierarchy, each observation j, for $j = 1, \dots, n_i$, is nested within a group i, for $i = 1, \dots, I$. Let $\boldsymbol{y}_i = (y_{1i}, \dots, y_{n_i i})$ be the n_i-dimensional response vector for observations in the i-th group. Conditionally on random effects denoted by $\boldsymbol{b}_i$, a GLMM assumes that the elements of $\boldsymbol{y}_i$ are independent, with density function from the exponential family, of the form

$$f_i(y_{ij}|\boldsymbol{b}_i) = \exp\left[\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} + c(y_{ij}, \phi)\right]$$

where $a(\cdot)$ and $c(\cdot)$ are specified functions, η_{ij} is the natural parameter and ϕ is the dispersion parameter. In addition, we have

$$\begin{aligned} E[y_{ij}|\boldsymbol{b}_i] &= a'(\eta_{ij}) = \mu_{ij} \\ Var[y_{ij}|\boldsymbol{b}_i] &= \phi\, a''(\eta_{ij}) \end{aligned}$$

A monotonic, differentiable link function $g(\cdot)$ specifies the function of the mean that the model equates with the systematic component. Usually, the canonical link function is used, i.e., $g = a'^{\,-1}$. From now on, without loss of generality, the canonical link function is used. In this case, the model is the following [31]:

$$\begin{aligned} \boldsymbol{\mu}_i &= E[\boldsymbol{Y}_i|\boldsymbol{b}_i] \qquad i = 1, \ldots, I \\ g(\boldsymbol{\mu}_i) &= \boldsymbol{\eta}_i \\ \boldsymbol{\eta}_i &= X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N_q(\boldsymbol{0}, \Psi) \qquad ind. \end{aligned} \tag{1}$$

where i is the group index, I is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{I} n_i = J$. $\boldsymbol{\eta}_i$ is the n_i-dimensional linear predictor vector. In addition, X_i is the $n_i \times (p+1)$ matrix of fixed-effects regressors of observations in group i, $\boldsymbol{\beta}$ is the $(p+1)$-dimensional vector of their coefficients, Z_i is the $n_i \times q$ matrix of regressors for the random effects, $\boldsymbol{b}_i$ is the $(q+1)$-dimensional vector of their coefficients and Ψ is the $q \times q$ within-group covariance matrix of the random effects. Fixed effects are identified by parameters associated with the entire population, whereas random ones are identified by group-specific parameters.

Our proposed generalised mixed-effects tree (GMET) method expands the use of tree-based mixed models to different classes of response variables from the exponential family. At the same time, the method can deal with the grouped data structure as GLMMs do. We now specify the GMET model. The random component of this model consists of a response variable Y from a distribution in the exponential family. The fixed part in the GMET is not linear as in (1), but is replaced by the function $\boldsymbol{f}(X_i)$ that is estimated through a tree-based algorithm. Thus, the matrix formulation of the model is the following:

$$\begin{aligned} \boldsymbol{\mu}_i &= E[\boldsymbol{Y}_i|\boldsymbol{b}_i] \qquad i = 1, \ldots, I \\ g(\boldsymbol{\mu}_i) &= \boldsymbol{\eta}_i \\ \boldsymbol{\eta}_i &= \boldsymbol{f}(X_i) + Z_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N_q(\boldsymbol{0}, \Psi) \qquad ind. \end{aligned} \tag{2}$$

where i is the group index, I is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{I} n_i = J$. In addition, $\boldsymbol{\eta}_i$ is the n_i-dimensional linear predictor vector and $g(\cdot)$ is the link function. Finally, X_i is the $n_i \times (p+1)$ matrix of fixed-effects regressors of observations in group i, Z_i is the $n_i \times q$ matrix of regressors for the random effects, $\boldsymbol{b}_i$ is the $(q+1)$-dimensional vector of their coefficients and Ψ is the $q \times q$ within-group covariance matrix of the random effects. As in a GLMM, $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i'}$ are independent for $i \neq i'$. Fixed effects are identified by a non-parametric CART tree model associated with the entire population, whereas random ones are identified by group-specific parameters.

Without loss of generality, let us now specify model (2) for the case of a binary random variable and univariate random effect. The logit function is the canonical link function:

$$g(\mu_{ij}) = g(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \text{logit}(p_{ij}).$$

Here, the random-effects structure simplifies to a random intercept. The model formulation for observation y_{ij} may therefore be written as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \qquad i = 1,\ldots,I \qquad j = 1,\ldots,n_i \\ p_{ij} &= E[Y_{ij}|b_i] \\ \text{logit}(p_{ij}) &= f(\boldsymbol{x}_{ij}) + b_i \\ b_i &\sim N(0,\psi) \qquad ind. \end{aligned} \tag{3}$$

where we observe $\boldsymbol{x}_{ij} = (x_{1ij},\ldots,x_{ijp})^T$, a $(p+1)$-dimensional vector of fixed-effects covariates for each observation j in group i.

2.2. Generalised Mixed-Effects Tree Estimation

In this subsection we show the algorithm for the estimation of the parameters of the GMET model (2). Following the approach of the RE-EM tree, the basic idea behind the algorithm is to disentangle the estimation of fixed and random effects, with the difference that the GMET algorithm is not iterative. The structure of the algorithm is the following:

1. Initialise the estimated random effects $\boldsymbol{b}_i$ to zero.
2. Estimate the target variable μ_{ij} through a generalised linear model (GLM), given fixed-effects covariates $\boldsymbol{x}_{ij} = (x_{ij1},\ldots,x_{ijp})^T$ for $i = 1,\ldots,I$ and $j = 1,\ldots,n_i$. Get estimate $\hat{\mu}_{ij}$ of target variable μ_{ij}.
3. Build a regression tree approximating f using $\hat{\mu}_{ij}$ as dependent variable and $\boldsymbol{x}_{ij} = (x_{ij1},\ldots,x_{ijp})^T$ as vector of covariates. This regression tree identifies a number L of terminal nodes R_ℓ, for $\ell = 1,\ldots,L$, and each observation ij, described by its set of covariates $\boldsymbol{x}_{ij}$, belongs to one of the terminal nodes. Through this regression tree, we define a set of indicator variables $I(\boldsymbol{x}_{ij} \in R_\ell)$, for $\ell = 1,\ldots,L$, where $I(\boldsymbol{x}_{ij} \in R_\ell)$ takes value 1 if observation ij belongs to the ℓ-th terminal node and 0 otherwise.
4. Fit the mixed effects model (2), using y_{ij} as a response variable and the set of indicator variables $I(\boldsymbol{x}_{ij} \in R_\ell)$ as fixed-effects covariates (dummy variables). Specifically, for $i = 1,\ldots,I$ and $j = 1,\ldots,n_i$, we have $g(\mu_{ij}) = I(\boldsymbol{x}_{ij} \in R_\ell)\gamma_\ell + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$. Extract $\hat{\boldsymbol{b}}_i$ from the estimated model.
5. Replace the predicted response at each terminal node R_ℓ of the tree with the estimated predicted response $g(\hat{\gamma}_\ell)$ from the mixed-effects model fitted in step 4.

The GLM in step 2 is fitted through maximum likelihood. The maximum likelihood estimates can be found using an iteratively reweighted least squares algorithm or a Newton–Raphson method [32].

The fitting of the tree in step 3 can be achieved using any tree algorithm, based on any tree-growing rules that are desired. Here, tree building is based on the CART tree algorithm [7]. After building a large tree T_0, pruning is advised to avoid overfitting on training data. In principle, any tree-pruning rule could be used; here, we propose cost-complexity pruning [33]. It considers a sequence of nested trees indexed by a nonnegative tuning parameter α which controls the trade-off between the subtree's complexity and its fit to the training data. For each value of α exists a subtree $T \subset T_0$ to minimise

$$\sum_{\ell=1}^{|T|} \sum_{x_i \in R_\ell} (y_i - \hat{y}_{R_\ell})^2 + \alpha|T|. \tag{4}$$

Here, $|T|$ indicates the number of terminal nodes of tree T. When $\alpha = 0$, then the subtree T will simply be equal to T_0. However, as α increases, the quantity (4) will tend to be minimised for a smaller subtree. We can select a value of α using a validation set or using k-fold cross-validation: for example, we can pick $\tilde{\alpha}$ to minimise the average CV error. Tree building and pruning is implemented in `R` library `rpart` [34], according to the CART tree-building algorithm and cost-complexity pruning. In order to ensure that initial trees are sufficiently large, we set the complexity parameter to zero. Thus, the largest tree is grown then pruned based on ten-fold cross-validation error. Instead of choosing the tree that achieves the lowest CV error, we use the so-called *1-SE rule*: any CV error within one standard error of the achieved minimum is marked as being equivalent to the minimum. Among all these equivalent models in terms of CV error, the simplest one is chosen as the final tree model.

The generalised linear mixed model in step 4 can be estimated using fitting techniques that were previously described. Different statistical packages can estimate those types of models: the `glmer` function of the `R` library `lme4` [35] is used here. It fits a generalised linear mixed model via maximum likelihood. For a GLMM the integral must be approximated: the most reliable approximation is the adaptive Gauss–Hermite quadrature, at present implemented only for models with a single scalar random effect; otherwise, Gaussian quadrature is used [36,37].

For what concerns the time efficiency, the GMET algorithm is very fast. Indeed, being a non-iterative algorithm, its running time is approximately equal to the sum of three steps' running times, i.e., the ones to fit a GLM (step 2), a regression tree (step 3) and a GLMM (step 4).

Predictions for New Observations

After estimating a GMET, it is possible to make out-of-sample predictions for new observations. Suppose the tree is estimated on data from groups $i = 1, \ldots, I$ for observations y_{ij}, $j = 1, \ldots, n_i$. Given a new observation $\boldsymbol{x}_{ij'}$, we are able to output its corresponding response, since we know the estimation of the fixed-effects function $f(\cdot)$, of the random effects $\boldsymbol{b}_i$ and of the associated covariance matrix Ψ. The algorithm is able to provide two types of prediction, depending on whether the group i to which the new observation $\boldsymbol{x}_{ij'}$ belongs is a new group (i.e., not observed in the data used to train the model) or not:

- Predict response $y_{ij'}$ given a new observation $\boldsymbol{x}_{ij'}$ for a group in the sample $i \in \{1, \ldots, I\}$. We define it a *group-level prediction*.
- Predict response $y_{i'j'}$ given an observation $\boldsymbol{x}_{i'j'}$ for a group i' for which there were no observations in our sample, or for which we do not know the relevant group. We define it a *population-level prediction*.

Following the classical approaches for prediction in mixed-effects models [8,38], for the first type of prediction, we estimate $f(\boldsymbol{x}_{ij'})$ using the estimated tree and attributes $\boldsymbol{x}_{ij'}$ and then add $\boldsymbol{z}_{ij'}^T \boldsymbol{b}_i$ on the linear predictor scale, and get back to the response scale through the inverse link function $g^{-1}(\cdot)$. As we underlined before, random-effects coefficients $\boldsymbol{b}_i$ are known from the estimation process. For the second type of prediction, since we have no information with which to evaluate $\boldsymbol{b}_i$, we set it to its expected value of 0, yielding the value $\hat{f}(\boldsymbol{x}_{i'j'})$, and transform it back to the response scale through the inverse link function. As noted in Sela and Simonoff [8], in this case we might expect that methods that do not incorporate random effects would have comparable performances to those that do, as long as the sample is large enough so that the fixed-effects function $f(\boldsymbol{x}_{ij'})$ is well-estimated by both types of methods.

3. Simulation Study

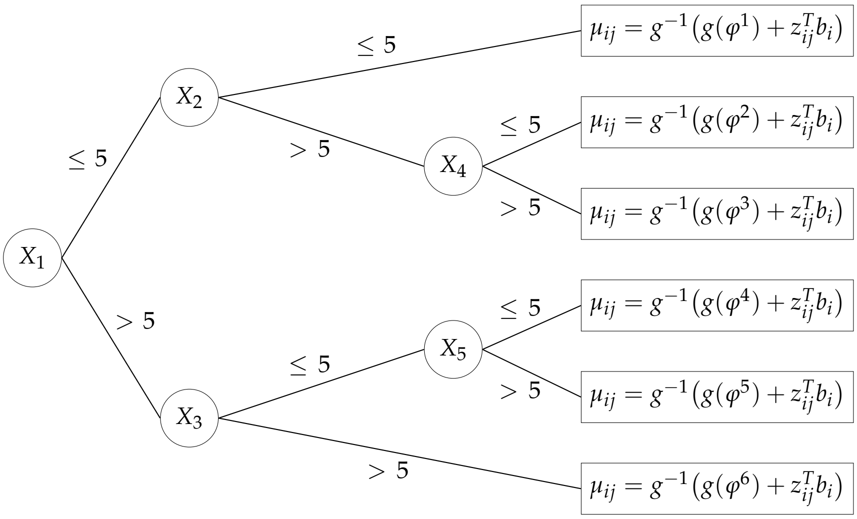
In this section we compare the performance of the proposed GMET method to the performances of standard classification trees and different types of mixed-effects models on simulated binary outcome datasets.

We used a variation of a simulation design proposed in Hajjem et al. [10] and followed the data generating process presented in their paper. We simulated a two-level data structure of $I = 50$ groups with $n_i = 60$ observations each: 10 observations in each group were included in the training sample, and the other 50 observations constituted the test sample. Therefore, $N_{\text{train}} = 500$ and $N_{\text{test}} = 2500$. By setting $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$, the response values y_{ij} were simulated according to a Bernoulli distribution with conditional probability of success μ_{ij}. Both fixed and random effects were used to generate μ_{ij}. Overall, we considered 10 different data generating processes (DGPs) outlined in Table 1 by combining different fixed and random-effect specifications[3].

Let us define the fixed-effect structure. Eight random variables $X_1, \ldots, X_8$, independent and uniformly distributed in the interval $[0, 10]$, were generated. While all of them were being used as predictors, only five of them were actually used to generate μ_{ij}, based on the tree rule summarised in Figure 1. Each observation was classified into one of the six terminal nodes according to the values $x_{ij1}, \ldots, x_{ij5}$. Within each leaf, values $\varphi^1, \ldots, \varphi^6$ denote the probabilities of success when the random effects b_i are equal to zero:

Leaf 1: if $x_{1ij} \leq 5 \wedge x_{2ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^1) + z_{ij}^T b_i\big)$;
Leaf 2: if $x_{1ij} \leq 5 \wedge x_{2ij} > 5 \wedge x_{4ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^2) + z_{ij}^T b_i\big)$;
Leaf 3: if $x_{1ij} \leq 5 \wedge x_{2ij} > 5 \wedge x_{4ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^3) + z_{ij}^T b_i\big)$;
Leaf 4: if $x_{1ij} > 5 \wedge x_{3ij} \leq 5 \wedge x_{5ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^4) + z_{ij}^T b_i\big)$;
Leaf 5: if $x_{1ij} > 5 \wedge x_{3ij} > 5 \wedge x_{5ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^5) + z_{ij}^T b_i\big)$;
Leaf 6: if $x_{1ij} > 5 \wedge x_{3ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^6) + z_{ij}^T b_i\big)$;

where $g(\cdot)$ is the logit link function. Two different possibilities were specified for the fixed effects: in the *large* fixed-effects specification, the standard deviation of the typical probabilities across the leaves was higher than in the *small* one (0.37 versus 0.24).



X_1
X_2
X_3
X_4
X_5
≤ 5
> 5
$\mu_{ij} = g^{-1}\big(g(\varphi^1) + z_{ij}^T b_i\big)$
$\mu_{ij} = g^{-1}\big(g(\varphi^2) + z_{ij}^T b_i\big)$
$\mu_{ij} = g^{-1}\big(g(\varphi^3) + z_{ij}^T b_i\big)$
$\mu_{ij} = g^{-1}\big(g(\varphi^4) + z_{ij}^T b_i\big)$
$\mu_{ij} = g^{-1}\big(g(\varphi^5) + z_{ij}^T b_i\big)$
$\mu_{ij} = g^{-1}\big(g(\varphi^6) + z_{ij}^T b_i\big)$

Figure 1. Mixed-effects tree structure used to generate the conditional probability of success μ_{ij} in the simulation study.

The random component $b_i \sim N(0, \Psi)$ was generated according to three different possibilities:

- No random effects: $\Psi = 0$;
- Random intercept: $z_{ij} = 1 \quad \forall i, \forall j$ and $\Psi = \psi_{11}$;
- Random intercept and slope, which add a linear random effect for the fixed-effect covariate X_1, uncorrelated from the random effect on the intercept. That is, $z_{ij} = [1 \quad x_{1ij}]^T \quad \forall i, \forall j$ and $\Psi = \begin{bmatrix} \psi_{11} & 0 \\ 0 & \psi_{22} \end{bmatrix}$.

Within each fixed-effects scenario with random effects, we considered two specifications (*low* and *high*) for the covariance matrix Ψ to account for different levels of magnitude of the between-group variability.

Table 1. Data generating processes (DGP) for the simulation study.

DGP	**Random component**				**Fixed component**						
	Structure	**Effect**	ψ_{11}	ψ_{22}	**Effect**	φ^1	φ^2	φ^3	φ^4	φ^5	φ^6
1	No random effect	–	–	–	Large	0.10	0.20	0.80	0.20	0.80	0.90
2		–			Small	0.20	0.40	0.70	0.30	0.60	0.80
3	Random Intercept	Low	4.00	–	Large	0.10	0.20	0.80	0.20	0.80	0.90
4		High	10.00	–							
5		Low	0.50	–	Small	0.20	0.40	0.70	0.30	0.60	0.80
6		High	4.00	–							
7	Random Intercept and Slope	Low	2.00	0.05	Large	0.10	0.20	0.80	0.20	0.80	0.90
8		High	5.00	0.25							
9		Low	0.25	0.01	Small	0.20	0.40	0.70	0.30	0.60	0.80
10		High	2.00	0.05							

Simulation Results

We ran eight different models for each one of the 10 DGPs:

- A standard binary classification tree model (*Std*);
- A random intercept GMET model (*RI*);
- A random intercept and slope GMET model (*RIS*);
- A parametric mixed-effects logistic regression model (*MElog*) that used the true model leaves' indicators as fixed covariates and the true random effect structure;
- A parametric mixed-effects logistic regression model (*GLMM*) that used $(x_1, \ldots, x_8)$ as fixed covariates and the true random effect structure;
- The *GLMERT* algorithm proposed in [11] considering $(x_1, \ldots, x_8)$ as fixed covariates and the true random effect structure;
- The *GMERT* algorithm proposed in [10] considering $(x_1, \ldots, x_8)$ as fixed covariates and the true random effect structure;
- The *BiMM* algorithm proposed in [12] considering $(x_1, \ldots, x_8)$ as fixed covariates and a random intercept[4].

As noted in Hajjem et al. [9], the *MElog* model could not be a real competitor of any other model. Indeed, it is not possible in practice to specify this parametric structure without knowing the underlying data generating process. This model only serves as a reference for the performances of the other models. In tree-based models, we fixed to 10 the maximum depth parameter and to 20 the minimum number of observations necessary to attempt a split[5]. After fitting each model on the training set, we could compute the corresponding predicted probability $\hat{\mu}_{ij}$ and the predicted class $\hat{y}_{ij}$ of observation j in group i in the test dataset. While the former was directly estimated by the algorithm, the latter depended on the threshold value μ_k^* used to classify subjects in the test set: $\hat{\mu}_{ij} \geq \mu_k^* \Rightarrow \hat{y}_{ij} = 1$ where $(i, j) \in$ test. There were at most K distinct fitted values μ_k, with $K \leq I|T|$. We used each of them to classify observations in the training set and we fixed the threshold μ_k^* as the one that yields the closest proportion of class 1 to the actual proportion of class 1 in the training set.

We measured the predictive performance by:

- The *predictive mean absolute deviation*

$$\text{PMAD} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{I} \sum_{j=1}^{n_i^{\text{test}}} |\mu_{ij} - \hat{\mu}_{ij}|$$

- The *predictive misclassification rate* (PMCR)

$$\text{PMCR} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{I} \sum_{j=1}^{n_i^{\text{test}}} |y_{ij} - \hat{y}_{ij}|.$$

The mean, median, standard deviation, minimum and maximum of the PMAD and the PMCR over 100 runs were calculated and are reported in Table 2.

We observed that when there was no random effect (DGPs 1 and 2), the standard classification tree algorithm performed better than the mixed-effects models, especially when the fixed effect was large. Nonetheless, in the latter scenario, the performances of *GLMERT* and *GMERT* were very close to *Std* ones, proving to be robust even in absence of a true random effect. However, when random effects were present (DGPs 3 to 10), mixed-effects classification trees performed better than the standard classification tree in terms of average PMAD and PMCR. *BiMM* is the only mixed-effects tree algorithm whose performance was very close to *Std* ones, for all DGPs[6]. When the DGP included only a random intercept, *GLMERT* had the best predictive performance, and was directly followed by *RI*. When the true random effect structure included both random intercept and random slope, *GMERT*, *GLMERT* and *RIS* performances were very close. There was a slightly better performance by *GLMERT* when the fixed effect was large and of *RIS* when the fixed effect was small. The highest improvement in PMAD using a mixed tree model was observed when both the fixed and the random effects were large. The lowest improvement was observed when both the fixed and the random effects were small. Analogous statements can be made about PMCR. In addition, GMET performed better than standard trees even when we fit a mixed tree whose random component was over-specified (as in DGPs 3–6, *Std vs RIS*) or under-specified (as in DGPs 7–10, *Std vs RI*) in relation to the true data generating process.

Table 2. Results of the 100 simulation runs in terms of predictive probability mean absolute deviation (PMAD) and predictive misclassification rate (PMCR) for the eight models for the 10 DGPs. DGPs for which the performance gap between MElog and GMET was the largest or the smallest are marked in bold.

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
1	NO RANDOM EFFECT	Large	Std	5.01	4.59	1.93	2.10	9.83	16.76	16.46	1.55	14.64	21.68
			RI	20.89	20.98	2.34	13.43	24.92	31.52	31.50	2.54	24.16	36.68
			RIS	20.91	21.02	2.22	13.18	25.21	31.12	31.30	2.14	23.20	35.96
			MElog	3.36	3.28	1.15	1.30	5.84	17.55	16.08	3.30	13.76	24.64
			GLMM	21.61	21.58	0.78	19.88	23.14	30.10	30.20	0.92	27.52	31.56
			GLMERT	5.73	5.43	2.17	2.37	11.02	19.38	18.50	3.01	14.76	25.04
			GMERT	4.85	4.33	1.84	1.96	9.45	17.80	17.70	1.73	15.12	21.64
			BiMM	21.54	23.09	3.23	16.63	26.21	30.49	30.52	1.33	25.16	33.44
2		Small	Std	9.97	10.22	3.29	4.49	17.62	32.24	32.72	2.39	28.00	38.64
			RI	13.66	13.58	1.82	10.48	18.13	37.24	37.42	2.11	32.68	41.48
			RIS	13.89	13.68	1.83	10.98	18.31	37.36	37.40	1.92	33.52	41.96
			MElog	4.07	4.02	1.35	1.42	7.74	28.84	28.80	1.79	25.96	34.48
			GLMM	15.43	15.35	0.54	14.09	16.67	37.44	37.48	1.20	34.72	40.08
			GLMERT	10.10	10.01	3.01	6.59	15.40	34.14	34.08	2.72	29.00	38.80
			GMERT	10.03	10.08	2.87	6.42	14.54	33.31	32.86	4.23	28.80	42.64
			BiMM	12.60	13.50	1.77	9.65	15.10	34.60	34.52	1.79	31.12	38.56

Table 2. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				mean	**median**	**sd**	**min**	**max**	**mean**	**median**	**sd**	**min**	**max**
3	INTERCEPT Low	Large	Std	23.39	22.95	2.93	18.01	29.90	29.26	28.62	3.31	23.40	36.36
			RI	**18.28**	18.12	1.57	13.81	22.98	**26.98**	26.96	2.05	21.92	32.20
			RIS	18.46	18.39	1.59	14.01	22.79	27.09	26.96	2.03	22.08	32.24
			MElog	**8.69**	8.61	0.75	7.60	10.85	**19.65**	19.46	1.12	17.72	23.24
			GLMM	18.62	18.67	1.05	16.74	20.96	26.69	26.70	1.56	23.80	30.40
			GLMERT	11.95	11.94	2.54	7.83	17.59	21.93	21.40	3.18	18.00	30.76
			GMERT	23.70	22.87	3.02	19.43	28.76	29.05	28.52	3.99	23.72	38.00
			BiMM	27.68	27.93	2.03	22.61	31.17	35.52	35.28	2.14	30.72	40.24
4	High		Std	31.70	31.94	2.58	26.22	36.78	36.23	36.20	3.12	30.32	44.16
			RI	15.38	15.46	1.51	11.96	18.57	20.68	20.68	1.97	16.64	25.76
			RIS	15.44	15.67	1.44	12.03	18.53	20.78	20.80	1.97	16.68	25.92
			MElog	8.21	8.06	0.95	6.21	11.01	15.78	15.70	1.50	12.56	20.12
			GLMM	15.40	15.36	1.29	12.01	17.76	20.66	20.60	1.93	16.40	24.88
			GLMERT	10.65	10.50	1.32	8.40	12.76	18.08	17.92	1.14	16.48	19.96
			GMERT	29.67	29.39	2.80	25.92	35.46	32.68	31.36	4.25	27.84	42.76
			BiMM	32.69	32.48	2.00	28.93	35.80	38.39	38.52	2.60	31.12	43.84
5	Low	Small	Std	15.79	15.87	2.39	10.13	22.90	34.30	34.92	2.35	29.00	38.56
			RI	15.68	15.77	1.68	13.11	19.26	35.74	35.74	2.30	31.24	43.12
			RIS	15.87	15.89	1.61	13.14	19.18	35.74	35.64	2.06	31.72	42.72
			MElog	8.55	8.61	0.92	6.45	10.73	28.80	28.66	0.99	25.84	30.96
			GLMM	16.48	16.35	0.59	15.13	18.23	36.47	36.60	1.21	33.52	39.32
			GLMERT	13.28	13.37	1.12	11.62	15.21	33.35	32.86	2.35	30.64	39.84
			GMERT	14.63	15.14	1.38	11.91	16.65	33.90	32.74	3.16	31.12	42.20
			BiMM	16.48	16.38	2.01	12.89	20.41	36.21	35.56	1.87	33.40	41.40
6	High		Std	27.98	28.16	2.33	23.28	32.46	41.23	40.88	3.09	35.92	50.44
			RI	**14.02**	13.99	1.62	10.01	17.45	**25.87**	26.14	2.41	20.64	30.56
			RIS	14.13	14.17	1.66	10.08	17.29	25.89	26.00	2.37	20.68	30.52
			MElog	**9.41**	9.43	1.10	7.24	11.79	**22.85**	23.22	1.66	20.00	26.36
			GLMM	14.24	14.13	1.05	11.95	16.82	25.98	25.88	2.02	22.32	30.96
			GLMERT	13.05	12.49	2.85	9.54	19.24	25.98	25.48	2.71	22.40	31.28
			GMERT	26.61	27.13	2.44	21.32	30.06	32.79	32.90	2.65	27.76	37.96
			BiMM	27.27	27.60	2.15	23.61	30.45	40.83	40.72	2.80	33.32	46.76
7	INTERCEPT & SLOPE Low	Large	Std	22.16	22.47	2.28	17.32	27.38	28.08	28.60	2.69	22.32	34.20
			RI	20.08	20.05	1.38	15.17	22.67	28.52	28.44	1.51	23.48	30.80
			RIS	**19.64**	19.67	1.29	16.00	22.64	**28.34**	28.14	1.44	24.20	30.68
			MElog	**9.95**	10.00	0.95	8.12	12.78	**20.09**	20.00	0.90	18.44	22.20
			GLMM	19.93	19.93	1.10	17.59	21.92	27.93	27.88	1.42	25.04	31.00
			GLMERT	12.10	11.80	1.57	10.30	15.72	21.76	21.92	1.04	20.24	24.32
			GMERT	14.71	14.89	1.34	12.55	16.85	23.05	22.62	1.35	21.68	25.76
			BiMM	26.39	26.53	1.50	22.72	28.52	35.04	35.40	1.80	31.00	38.52
8	High		Std	32.57	32.42	2.85	26.92	38.29	37.46	36.82	4.12	30.36	49.68
			RI	17.29	17.38	1.53	13.56	20.87	21.66	21.42	2.19	17.68	25.64
			RIS	15.82	15.89	1.56	11.80	18.42	20.72	20.58	2.18	17.08	24.72
			MElog	9.50	9.48	0.82	7.72	10.97	16.09	16.16	1.40	12.24	19.00
			GLMM	15.87	15.75	1.34	13.55	18.82	20.57	20.28	2.14	16.52	25.60
			GLMERT	13.08	13.38	1.62	10.15	15.57	18.92	19.54	1.64	16.08	20.76
			GMERT	17.63	17.38	1.34	16.04	20.71	21.33	21.66	2.05	18.08	25.04
			BiMM	33.62	33.40	1.61	30.70	37.02	39.41	39.48	2.71	33.48	44.80

Table 2. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
9	Low	Small	Std	16.55	16.78	2.25	11.52	20.62	35.13	35.12	2.43	29.76	39.52
			RI	15.94	15.62	1.43	12.37	18.89	36.37	36.18	2.11	31.92	41.04
			RIS	15.83	15.55	1.47	12.19	18.91	36.17	36.28	1.88	31.92	41.08
			MElog	9.04	8.84	0.87	7.35	11.39	29.03	29.06	0.97	26.72	31.20
			GLMM	16.81	16.66	0.76	15.11	18.64	36.71	36.72	1.36	34.00	40.20
			GLMERT	13.45	13.64	2.06	10.04	17.46	32.92	32.72	2.93	28.36	38.64
			GMERT	13.05	13.04	1.89	10.38	16.16	32.81	33.04	2.59	28.68	37.12
			BiMM	16.37	15.82	1.72	13.96	19.86	36.47	35.96	2.20	32.64	41.48
10	High		Std	26.95	26.57	2.26	22.70	31.94	40.45	39.98	3.19	33.52	47.76
			RI	15.76	15.90	1.40	12.71	18.94	27.97	27.90	2.19	22.52	32.76
			RIS	**15.28**	15.14	1.39	12.73	18.65	**27.61**	27.56	2.23	22.72	31.56
			MElog	**10.80**	10.76	1.10	7.86	13.74	**24.25**	24.24	1.75	20.48	28.16
			GLMM	15.45	15.43	1.00	13.18	17.42	27.65	27.88	2.08	23.12	31.96
			GLMERT	15.77	16.32	1.79	13.08	18.61	28.03	28.48	2.05	23.92	30.80
			GMERT	17.77	18.44	1.79	14.72	20.49	29.83	29.80	2.17	25.56	33.52
			BiMM	25.41	24.92	2.14	21.90	29.44	39.33	38.92	2.73	34.56	45.36

Next, we compare the performance of the GMET approach to the results of the MElog reference model. If the DGP did not include random effect, the difference between PMAD and PMCR was higher when the fixed effect was large (DGP 1). When the random effect was large and the fixed effect was small (DGPs 6 and 10), the GMET model performed similiarly to the MElog model. In terms of PMAD, the differences were 4.61% and 4.48% for DGPs 6 and 10, respectively; in terms of PMCR, they were 3.02% and 3.36%, respectively. The difference in predictive accuracy between the two models reached its maximum when the random effect was small and the fixed effect was large (DGPs 3 and 7). In terms of PMAD, the differences were 9.59% and 9.69% in DGPs 3 and 7, respectively; in terms of PMCR, they were 7.33% and 8.25%, respectively.

With respect to the other existing tree-based mixed-effects models, the fact that the GMET algorithm is not iterative makes it less performant when fixed and when its random effects are small and easier to be confused; and its performs better when they are large and easier to be disentangled. Moreover, its step through a *glm* makes it perform worse when the DGP includes only a large (in this case, nonlinear) fixed effect, but makes it competitive with the other existing methods when data have an important random-effects structure. In order to investigate the performance and to deepen the comparison across methods under different settings, we report, in Appendix B, additional simulations and results: we provide more details about the model's predictive quality in this simulation, e.g., the recovery of the right tree structure or the identification of the right number of leaves. We ran new simulations for different DGPs (linear and non-linear fixed-effects) and for a different response variable in the exponential family, i.e., Poisson. Results show that *GMET* on average outperformed all other tree-based methods when data had a linear structure, for both a binary and a Poisson response variable.

4. Case Study: Application of the Mixed-Effects Tree Algorithm to Education PoliMi Data

In this section, we describe the PoliMi dataset. We applied the generalised mixed-effects tree algorithm to these data. Using a GMET model, we could identify discriminating fixed-effects covariates and estimate the degree programme's effect on the predicted success probability. In addition, we also analysed the accuracy of this model for predicting dropout.

The PoliMi dataset consists of 18,612 records in Bachelor of Science (BSc) students that began between A.Y. 2010/2011 and 2013/2014. Students are nested within $I = 19$ degree programmes. Table 3 reports the list of the 19 degree programmes and the number of students enrolled in each degree program. A descriptive analysis showed that a high percentage of students leave the Politecnico before obtaining a degree. In particular, the sample shows a 37.11% dropout rate. Therefore, our goal was to find out which student-level indicators could discriminate between two different profiles: *dropout* and *graduate* students.

Table 3. Number of students enrolled in the 19 PoliMi degree programmes between A.Y. 2010/2011 and 2013/2014.

Degree Program	Number of Students
Aerospace Engineering	1127
Automation Engineering	538
Biomedical Engineering	1456
Building Engineering	671
Chemical Engineering	715
Civil and Environmental Engineering	405
Civil Engineering	855
Electrical Engineering	575
Electronic Engineering	567
Energy Engineering	1485
Engineering of Computing Systems	2173
Environmental and Land Planning Engineering	590
Industrial Production Engineering	288
Management Engineering	2750
Materials and Nanotechnology Engineering	637
Mathematical Engineering	575
Mechanical Engineering	2364
Physics Engineering	469
Telecommunications Engineering	372

We assumed a binary GMET model (3) where student j was nested within degree programme i. The response variable Y was the `status`, a two-level factor we coded as a binary variable:

- `status` $= 1$ for studies definitely completed with graduation;
- `status` $= 0$ for studies definitely concluded with dropping out.

We would like to make predictions at the very early stages of students' academic careers. Thus, we chose as predictors five variables available at the time of enrolment and three more variables collected just after the first semester of study. The list and explanation of student-level variables to be included as covariates is reported in Table 4. In addition, we chose as the grouping variable the degree programme at the time of enrolment (factor `DegreeProgramme`) which has 19 levels. The influence of the grouping factor on the predictor was modelled through a group-level intercept b_i. We randomly split the dataset into training and test subsets, with a ratio of 80% for training and 20% for evaluation. Thus, the training subset included 14,890 students and the test subset had 3722.

Table 4. A list and explanations of variables at the student level which were included as covariates in the GMET model.

Variable	**Description**	**Type of Variable**
`Sex`	gender	factor (2 levels: M, F)
`Nationality`	nationality	factor (Italian, foreigner)
`PreviousStudies`	high school studies	factor (*Liceo Scientifico*, *Istituto Tecnico*, Other)
`AdmissionScore`	PoliMi admission test result	real number
`AccessToStudiesAge`	age at the beginning of the BSc studies at PoliMi	natural number
`WeightedAvgEval1.1`	weighted average of the evaluations during the first semester of the first year	real number
`AvgAttempts1.1`	average number of attempts to be evaluated on subjects during the first semester of the first year (passed and failed exams)	real number
`TotalCredits1.1`	number of ECTS credits obtained by the student during the first semester of the first year	natural number

While growing the tree, we fixed to 10 the maximum depth parameter and to 20 the minimum number of observations necessary to attempt a split. Figure 2 shows the estimated mixed-effects tree for the probability of graduation. Every internal node had a corresponding condition that split it into two children: if the condition was true, observations were sent down the tree through the left child; if the condition was false, through the right child. In addition, all nodes reported two values: the estimated probability of graduation and the percentage of observations in the node over the total training set. We remind the reader that variable `PreviousStudies` has been coded as a three-level factor with levels S (*Liceo Scientifico*), T (*Istituto Tecnico*) and O (other high school studies). The number of ECTS obtained in the first semester of the first year was used as the first split: students who obtained less than 13 ECTS were associated with lower success probability (0.16 versus 0.86). Then, students were further classified using other explanatory variables: we can see that Italian students who obtained more than 24 ECTS had the highest predicted success probability (0.95). Other variables actually used to split smaller internal nodes were `Nationality` and `PreviousStudies`: in these nodes, students who attended *Istituto Tecnico* and foreign students had lower predicted success than the others. Through this model, it was possible to find out significant interactions among the covariates: for example, variable `Nationality` was used to split the group of students that obtained at least 13 ECTS, but this same variable did not appear in the complementary branch of the tree. Finally, covariates `Sex`, `AdmissionScore` and `AvgAttempts1.1` were not compared in the trees, so they do not appear to have strong influences on how one's studies end.

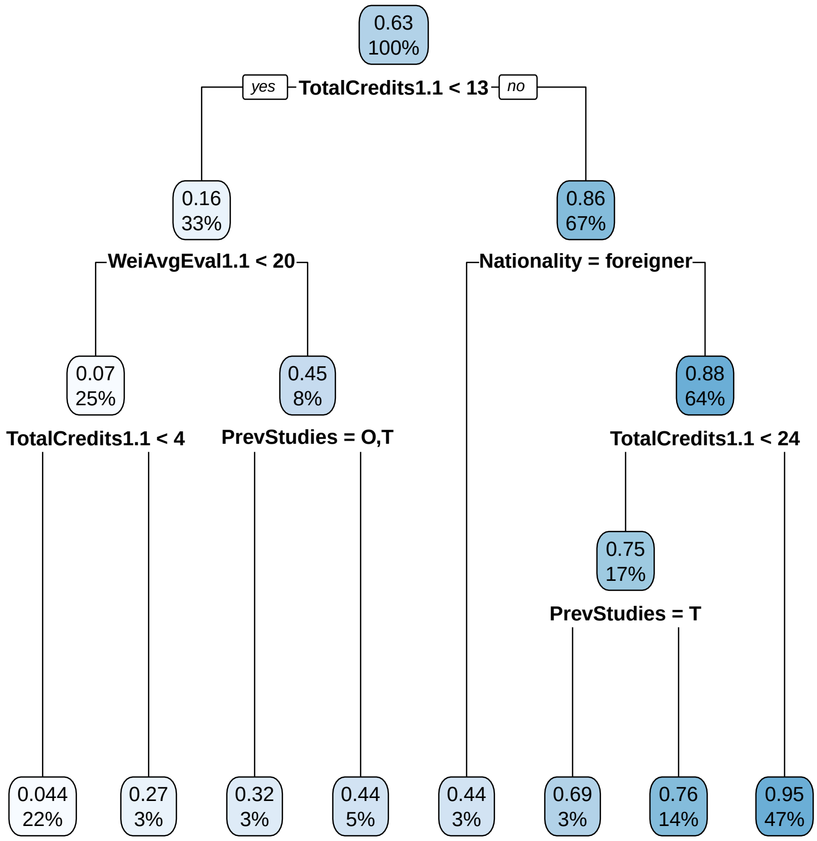


Figure 2. The estimated mixed-effects tree of model (3) for the probability of graduation. Each node reports the percentage of observations belonging to the node (second line of the node) and the estimated probability that responses relative to these observations are equal to 1 (first line of the node). Regarding the splitting criteria, left branches correspond to the case in which the condition is satisfied, and right branches correspond to the complementary case.

Using the tree structure in Figure 2, we could get population-level predictions for new observations that did not include the effect of the programme. However, if we also specified the level of the random effect covariate, our model was able to adjust this prediction to account for the effect and make a group-specific prediction. Indeed, we extracted coefficients $\hat{b}_i$ from the full estimated mixed model (3) and provide different predictions for different programmes within each leaf of the tree structure. Figure 3 shows the ranking of the 19 estimated random-effects intercepts, one for each degree program. Light blue points correspond to the point estimates $\hat{b}_i$, for $i = 1, \dots, 19$, and the horizontal black lines represent the 95% confidence intervals of the estimates. When the 95% confidence interval does not overlap with 0 (identified by the dashed vertical line), we have evidence to assert that the degree program's effect was significantly different from zero, i.e., from the average. For many groups, the 95% confidence interval does not overlap with the vertical line at zero, underlining substantial differences between the groups. If we use this model to estimate the probability of graduation, many degree programs will give results significantly different from the average. In particular, degree programs whose confidence intervals are entirely higher (lower) than zero are associated with higher (lower) dropout likelihood with respect to the average, all else being equal. After fixing all other covariates, *Environmental and Land Planning Engineering* and *Civil and Environmental Engineering* had large positive effects on the intercept: one student from one of those programmes improves the log odds by 1.051 or 0.705, respectively. On the contrary, studying either *Civil Engineering* or *Electrical Engineering* penalises the log odds by 0.680 and 0.546 respectively.

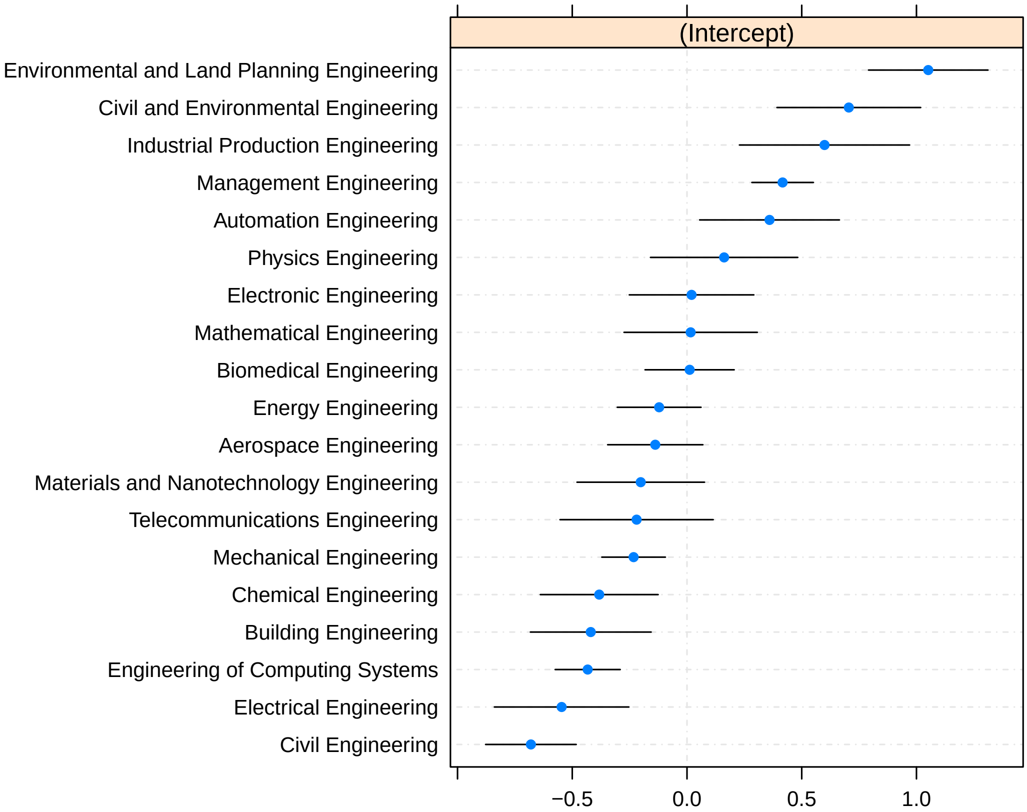


Figure 3. Estimated random intercept for each degree programme in model (3). For each engineering programme, the blue dot and the horizontal line mark the estimate and the 95% confidence interval of the corresponding random intercept.

Since we were using a multilevel model, we were able to account for the interdependence of observations by partitioning the total variance into different components due to the clustered data structure in model (3). The variance partition coefficient (VPC) is a possible measure of intraclass correlation: it is equal to the percentage of variation that is found at the higher level of hierarchy over the total variance [39]. The idea of VPC was extended using the latent variable approach, to define a method to partition the total variance in the case of a binary response and the group-specific intercept for the random-effects structure [40]. In this case, the variance partition coefficient was constant across all individuals, and it can be estimated as:

$$\text{VPC} = \frac{\hat{\psi}}{\hat{\psi} + \sigma^2_{lat}} = \frac{0.2988}{0.2988 + \pi^2/3} = 0.0612$$

where $\hat{\psi}$ is the estimated variance of the random intercept and σ^2_{lat} is the residual variability that can be explained by neither fixed effects, nor the group features that are represented by the random intercept. In this case, it is equal to the variance of the standard logistic distribution. This VPC value means that 6.12% of variation in the response is attributable to the classification by degree type. This value underlines the need to use a mixed model.

We can now evaluate the performance of the model and its predictive quality using the area under the ROC curve (AUC) and other performance indexes: accuracy, sensitivity and specificity. For each test observation, we were given a full set of covariates; therefore, we were able to compute an estimate $\hat{p}$ of the probability of successfully concluding a BSc and getting a degree. We used this estimate to define a binary classifier based on model (3): we chose $p_0 = 0.6$ as the optimal cutoff value through ROC curve analysis, as shown in Figure 4. For 20 iterations, we randomly split the observations into training and test sets. We fit a GMET model with the training set, and we classified test observations using the optimal threshold value. Finally, we computed the average accuracy, sensitivity and specificity values and their standard deviations, which are reported in Table 5. High values of accuracy, sensitivity and specificity indicate a good model. The model's performance was robust, as highlighted by the low standard deviations of the mean performance indexes and the high AUC, equal to 0.9127 (Figure 4). In addition, Table 6 reports the means and standard deviations of accuracy, sensitivity and specificity, computed separately for each degree program.

Table 5. Performance indexes of the classifier based on the mixed-effects tree of model (3).

Index	Mean	Std Deviation
Accuracy	0.860	0.006
Sensitivity	0.816	0.012
Specificity	0.886	0.008

Table 6. Performance indexes of the classifier based on the mixed-effects tree of model (3), computed for each degree program.

Degree Program	Accuracy Mean (sd)	Sensitivity Mean (sd)	Specificity Mean (sd)
Aerospace Engineering	0.880 (0.028)	0.845 (0.038)	0.897 (0.034)
Automation Engineering	0.880 (0.053)	0.798 (0.098)	0.925 (0.045)
Biomedical Engineering	0.894 (0.019)	0.860 (0.042)	0.912 (0.024)
Building Engineering	0.856 (0.042)	0.860 (0.080)	0.852 (0.050)
Chemical Engineering	0.877 (0.036)	0.889 (0.056)	0.873 (0.049)
Civil and Environmental Engineering	0.879 (0.038)	0.841 (0.081)	0.907 (0.052)
Civil Engineering	0.718 (0.041)	0.650 (0.044)	0.837 (0.060)
Electrical Engineering	0.849 (0.040)	0.840 (0.056)	0.867 (0.058)
Electronic Engineering	0.854 (0.037)	0.806 (0.078)	0.886 (0.053)
Energy Engineering	0.898 (0.023)	0.884 (0.059)	0.903 (0.022)
Engineering of Computing Systems	0.823 (0.022)	0.846 (0.029)	0.805 (0.030)
Environmental and Land Planning Engineering	0.851 (0.034)	0.782 (0.100)	0.878 (0.052)
Industrial Production Engineering	0.822 (0.091)	0.692 (0.164)	0.916 (0.068)
Management Engineering	0.873 (0.014)	0.765 (0.040)	0.931 (0.015)
Materials and Nanotechnology Engineering	0.907 (0.034)	0.867 (0.088)	0.918 (0.031)
Mathematical Engineering	0.893 (0.031)	0.851 (0.058)	0.908 (0.040)
Mechanical Engineering	0.863 (0.023)	0.841 (0.032)	0.875 (0.028)
Physics Engineering	0.902 (0.031)	0.852 (0.062)	0.930 (0.041)
Telecommunications Engineering	0.853 (0.058)	0.845 (0.087)	0.858 (0.061)

It is interesting to compare these average performance indexes against those obtained using different methods. Our approach had similar accuracy to a standard classification tree (0.878 versus 0.879), but its accuracy showed less variability across the iterations. For example, its standard deviation of accuracy was 0.5%; compare that to 2.8% for the classification tree. Since we were interested in the detection of dropout careers, we compared mean sensitivity using different models. Using mixed-effects trees, we attained higher sensitivity than using standard classification trees (0.835 versus 0.800). Thus, the choice of a mixed-effects model seemed appropriate: the degree programme is a meaningful covariate for the prediction of `status`. The mixed-effects tree was slightly less sensitive than a classifier built through a GLMM (0.835 versus 0.850), suggesting that a tree-like structure for fixed effects might not be as suitable as the GLMM one. However, it has other advantages, such as offering an easily interpretable model that could be graphically displayed and understood. Overall, the good performance of GMET in this application was due to two reasons. The first is that the variability at the highest level of grouping, i.e., degree programs, was not negligible, and therefore, taking it into account improved the predictive performance of the models. The second is that the good performance of the GLMM suggests that the association between the most important covariate, i.e., *TotaleCredits1.1* (the most relevant variable in the tree of Figure 2), and the response can be well approximated by a linear function. Therefore, $\hat{\mu}_{ij}$, estimated at step 2 of the GMET algorithm and used as the input for the tree built at step 3, was very precise and representative of the real dynamics, helping the GMET algorithm to fit the data well.

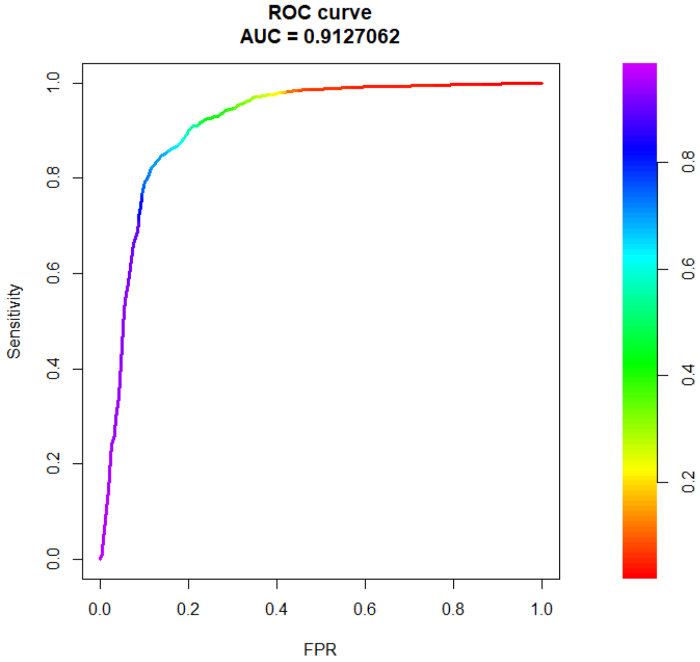


Figure 4. ROC curve computed on the PoliMi test set. Standing on this evidence, we chose 0.6 as the optimal value of p_0 to be used in the prediction as the threshold value for classification.

Appendix A reports the results of the application of GMERT and BiMM algorithms to the PoliMi case study and a comparison with GMET results presented in this section.

5. Conclusions

We proposed a multilevel tree-based model for a non-Gaussian response (GMET algorithm), showed a simulation study and applied the GMET algorithm to the PoliMi careers dataset as a tool to find student-level variables to discriminate between two different student profiles (graduate and dropout) and to estimate the degree programme's effect on the predicted success probability.

The GMET model can deal with a grouped data structure, while providing easily interpretable models that can outline complex interactions among the input variables. In the simulation study, the performance of the proposed mixed-effects tree method was a marked improvement over the CART model when the data generating process (DGP) included random effects, even if they were of small magnitude. In addition, the performance of the GMET model was similar to that of the benchmark logistic model that was fitted assuming the whole specification of the DGP. GMET's performance was comparable to that of other existing tree-based mixed-effects models, outperforming them when data had a linear structure, and it had a clear advantage in convergence time. Although our study focused on the binary response case, the mixed-effects tree approach could be extended to other types of response variables. Using a suitable link function, we could study if the method is appropriate to model different outcomes, such as count data or a multinomial factor response. Overall, the main advantages of the GMET algorithm are its flexibility and interpretability [41]. By relaxing the linear assumption of the fixed-effects part, the method could model more complex functional forms, easily treating potential interactions among covariates. This complexity is then summarised in a tree structure, which is easy to interpret and communicate. At the same time, when data present a hierarchy, the method is able to take into account the dependence structure within observations and to model it. In the educational data mining context, this aspect is essential in order to better understand students and the settings in which they learn. On the other hand, GMET, as CARTs, suffers from high variance. This means that if we split the training data into two parts at random, and fit a decision tree to both halves, the results could be quite different. Ensemble methods which use a mixed-effects tree as a base learner together with a random forest approach may be developed.

In our case study, the effectiveness of the GMET model in dropout prediction was comparable to the effectiveness of more established classification methods. A GMET model with high accuracy and sensitivity was obtained by considering information available at the time of the admission and the results of the first semester of studies. In addition, our work identifies a significant effect of the engineering programme on dropout probability. The estimated student success probability might be used as a tool to conduct policy experiments at the institutional level, aimed at identifying the best practices to help and retain at-risk students. In this setting, PoliMi started an experimental early intervening program that invites at-risk students (identified by the GMET algorithm) to attend dedicated tutorship to support them during the beginning of their studies at PoliMi.

In the context of the SPEET project, a future development could be the extension of our analysis to the other project partners in order to compare the programme effect at the country level. This would allow us to relate this effect to programme-level variables, and we could establish whether the same profiles of students at risk of dropout arise at country level. Moreover, in accordance with the validity and the potential of the GMET method when applied to modelling student dropout prediction, our future perspective goes in the direction of major applications in the learning analytics area. This method, when applied to educational data, can be a useful tool to support the definition of best practices and new tutoring programmes aimed at enhancing student performances and reducing student dropout. A worthwhile consideration is also the approach that teachers and students have with respect to its results. Indeed, this method is also valuable from the perspective of recommendation systems, since, if its results are interpreted and communicated in the right way, they can be used to drive students in their career choices.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/data6070074/s1, DataS1: The R code for the GMET algorithm.

Author Contributions: Conceptualization, F.I. and A.M.P.; Data curation, L.F.; Methodology, L.F., C.M. and A.M.P.; Resources, A.M.P.; Supervision, F.I. and A.M.P.; Writing—original draft, L.F.; Writing—review & editing, C.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in Supplementary Materials.

Acknowledgments: This work was within the Student Profile for Enhancing Engineering Tutoring (SPEET) project, funded by Erasmus$^+$. The authors are grateful to Umberto Spagnolini and Aldo Torrebruno for their comments and support during this work.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Application of GMERT Algorithm to PoliMi Case Study and Comparison with GMET Results

In this section, we describe the application of the GMERT algorithm proposed in [10] and the BiMM algorithm proposed in [12] to our case study on PoliMi SPEET data, to compare their results with our GMET ones (reported in Section 4).

We ran GMERT and BiMM algorithms considering the same set of fixed-effects covariates shown in Table 4 and a random intercept given by the grouping of students within degree programmes. Equivalently to GMET inputs, we fixed to 10 the maximum depth parameter and to 20 the minimum number of observations necessary to attempt a split in the GMERF algorithm. Since the BiMM algorithm does not receive in input *rpart* control parameters, we ran the algorithm with the default parameters. Figures A1 and A2 report the fixed-effects trees and the random intercepts estimated by GMERT and BiMM. Regarding the fixed-effects, the trees identified by GMET and GMERT are very similar: the variables that were determined important are coherent across the two methods (i.e., `TotalCredits1.1` as the most important one, followed by `WeiAvgEval1.1`, `PrevStudies` and `AccessAge`). BiMM tree performs a unique split, identifying `TotalCredits1.1` as the most important covariate. Regarding the random-effects, comparing Figures 3 and A2, we can observe that the random intercepts estimated by the three methods are quite consistent. In particular, the correlation coefficient between random intercepts estimated by GMET and GMERT is equal to 0.95, whereas the one between random intercepts estimated by GMET and BiMM is equal to 0.73. The variance of random intercepts ψ estimated by GMERT is smaller that estimated by GMET. Indeed, the VPC estimated by model GMERT is 0.0479 (against $VPC_{GMET} = 0.0612$). The variance ψ estimated by the BiMM algorithm is higher, and $VPC_{BiMM} = 0.0634$.

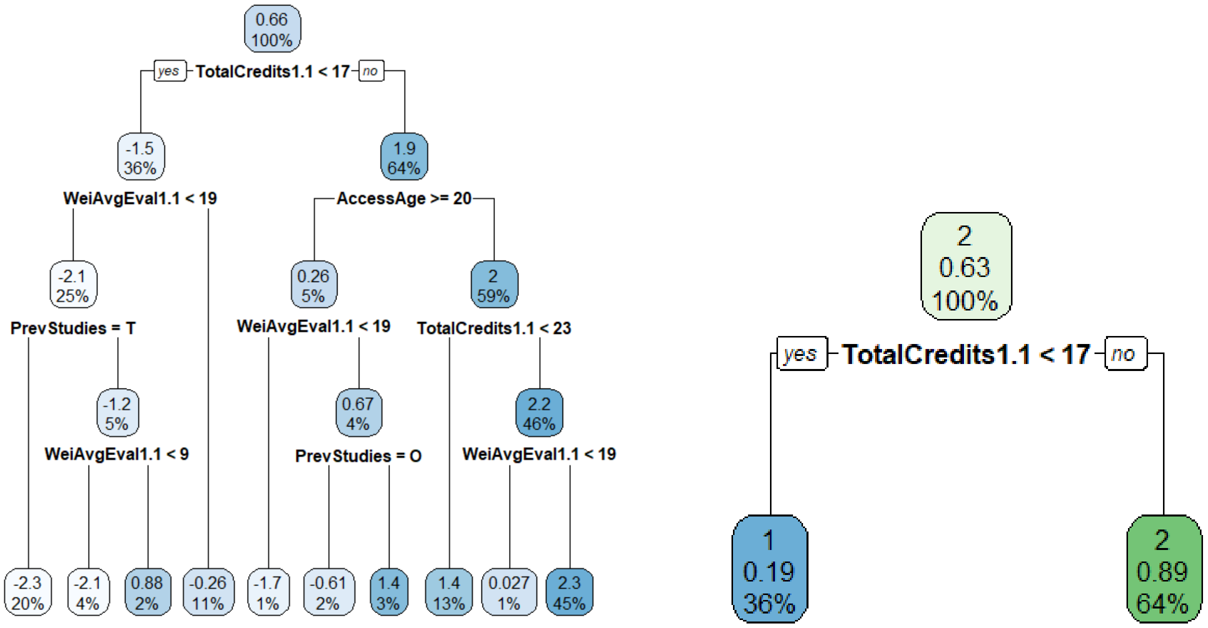


Figure A1. Fixed-effects trees estimated by the GMERT algorithm (**left panel**) and the BiMM algorithm (**right panel**) for the probability of graduation. GMERT tree leaves do not report probability of class 1 as GMET and BiMM leaves do, but they report the estimated linearised response variable (obtained using a first-order Taylor-series expansion). BiMM notation is 2 for graduate and 1 for dropout.

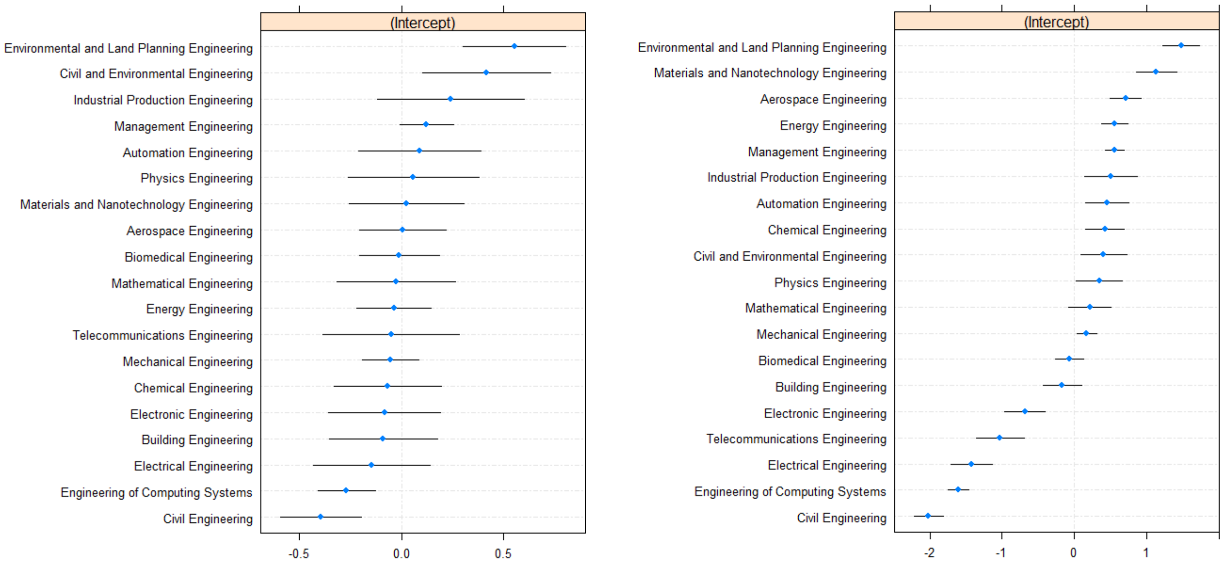


Figure A2. Random intercept for each degree programme, estimated by GMERT (**left panel**) and BiMM (**right panel**). For each engineering programme, the blue dot and the horizontal line mark the estimate and the 95% confidence interval of the corresponding random intercept.

Regarding the predictive performances of GMERT and BiMM, Figure A3 reports the ROC curves obtained on the test set and Table A1 reports performance indexes of the classifiers based on the two methods, computed following the same procedure of GMET. The predictive performances of GMET and GMERT were very similar, with the small differences that the AUC of GMERT was slightly higher than that of GMET, but the accuracy, sensitivity and specificity indexes of GMERT had higher values than those of GMET. It is worth noting that the time of convergence for GMERT was significantly higher than that for GMET. BiMM seemed to perform slightly worse than the other two methods in terms of predictive power.

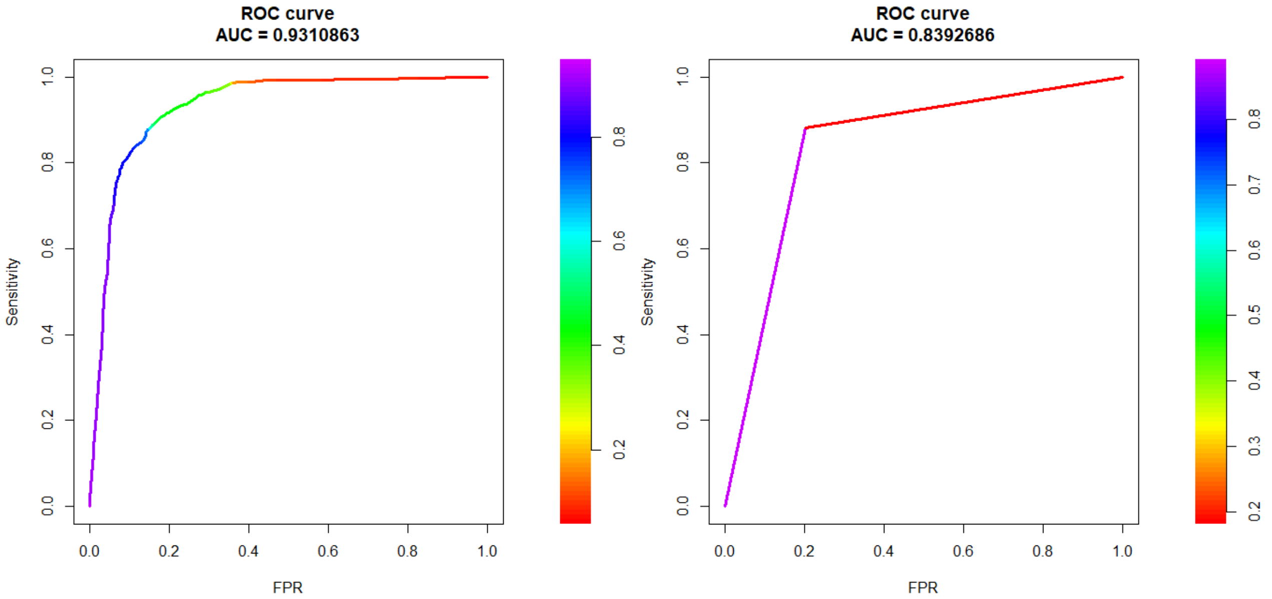


Figure A3. ROC curve computed on the PoliMi test set for the GMERT model (**left panel**) and BiMM model (**right panel**), respectively. Standing on this evidence, we choose 0.6 as the optimal value of p_0 to be used in the prediction as the threshold value for classification.

Table A1. Performance indexes of a classifier based on the mixed-effects tree estimated by GMERT and BiMM algorithms, computed on 20 iterations, randomly splitting the observations into training and test sets.

	GMERT		**BiMM**	
Index	**Mean**	**Std Deviation**	**Mean**	**Std Deviation**
Accuracy	0.861	0.008	0.849	0.012
Sensitivity	0.818	0.021	0.806	0.023
Specificity	0.891	0.013	0.874	0.015

Appendix B. Additional Simulations and Results

In this section, we provide more details about the simulations presented in Section 3 and also the results from other simulations with different DGPs.

Appendix B.1. Recovery of the Right Tree Structure

The predictive performances of the GMET algorithm and other tested methods are given in Table 2, Section 3. Here we present the results about the ability of the methods to recover the right tree structure. Following the approach presented in [10], three different ways of looking at this aspect are presented. We evaluate if the tree has: (1) the right number of leaves (i.e., six), (2) the right structure and right splitting covariates and (3) the right structure, right splitting covariates and right cutpoints. The third criterion was achieved if the estimated cutpoints came within 1 unit of the true ones (which were all 5), that is, if $4 < \text{cutpoint} < 6$ for all cutpoints. These criteria are in increasing order of difficulty. If the estimated tree achieved (3), then it achieved (2) and (1), and so on. Table A2 presents the results. In terms of number of leaves, *Std*, *GLMERT* and *GMERT* did fairly well with median numbers of leaves of six or sometimes five in the first four DGPs. *GLMERT* and *GMERT* tended to underestimate the number of leaves in the last six DGPs. *BiMM* tended to underestimate the number of leaves in all DGPs. Contrarily to the others, *GMET* tended to overestimate the number of leaves of the tree in all the DGPs. This result might depend on the second step of the GMET algorithm, in which the response variable is linearised through a GLM. As a consequence, the tree is built not using the original binary response $y \sim Be(p)$ as the target variable, but the $\hat{p}$ estimated by the GLM. This can lead to a different tree structure.

Table A2. Results of the 100 simulation runs, presented in Table 2, Section 3, in terms of recovering the right tree structure. # *right splits* reports the number of times out of 100 in which we obtained a tree of six leaves with the right splitting covariates; # *right cutpoints* reports the number of times out of 100 in which we obtained a tree of six leaves with the right splitting covariates and cutpoints (i.e., 4 < cutpoint < 5).

DGP	**Random Effect**	**Fixed Effect**	**Fitted Model**	**Number of Leaves**					**Right Tree Structure**	
				Mean	**Median**	**SD**	**Min**	**Max**	**# Right Splits**	**# Right Cutpoints**
1	NO RANDOM EFFECT	Large	Std	6.11	6.00	0.39	6.00	8.00	84	78
			RI	10.08	10.00	1.73	6.00	13.00	0	0
			RIS	10.11	10.00	1.78	7.00	14.00	0	0
			GLMERT	6.39	6.00	0.59	6.00	8.00	94	63
			GMERT	6.13	6.00	0.66	6.00	10.00	89	61
			BiMM	2.74	3.00	0.69	2.00	5.00	0	0
2		Small	Std	7.21	6.00	3.07	3.00	15.00	24	16
			RI	10.58	10.00	1.48	8.00	13.00	0	0
			RIS	10.58	10.00	1.43	8.00	14.00	0	0
			GLMERT	4.84	5.00	0.86	4.00	8.00	24	8
			GMERT	4.76	5.00	1.05	3.00	7.00	48	25
			BiMM	3.66	4.00	0.75	2.00	5.00	0	0

Table A2. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	Number of Leaves					Right Tree Structure	
				Mean	Median	SD	Min	Max	# Right Splits	# Right Cutpoints
3	INTERCEPT Low	Large	Std	7.24	6.00	2.14	4.00	14.00	31	21
			RI	10.32	11.00	1.97	6.00	13.00	2	0
			RIS	10.24	11.00	1.97	7.00	14.00	0	0
			GLMERT	6.24	6.00	0.63	5.00	8.00	75	60
			GMERT	5.95	6.00	1.09	3.00	9.00	87	68
			BiMM	3.11	3.00	0.83	2.00	5.00	0	0
4	High		Std	6.26	6.00	3.29	1.00	14.00	8	3
			RI	10.11	10.50	1.98	5.00	14.00	3	1
			RIS	10.08	10.00	1.68	6.00	13.00	0	0
			GLMERT	5.53	6.00	1.16	3.00	8.00	44	21
			GMERT	4.45	5.00	1.80	1.00	8.00	45	26
			BiMM	3.18	3.00	0.56	2.00	4.00	0	0
5	Low	Small	Std	7.32	6.00	3.62	3.00	17.00	8	5
			RI	10.18	10.00	1.54	6.00	14.00	0	0
			RIS	10.29	10.00	1.71	6.00	13.00	0	0
			GLMERT	4.79	5.00	0.84	4.00	7.00	10	3
			GMERT	4.76	4.50	1.57	2.00	10.00	36	12
			BiMM	3.66	4.00	0.71	3.00	5.00	0	0
6	High		Std	5.82	4.00	3.75	2.00	16.00	0	0
			RI	10.03	10.00	1.95	6.00	15.00	0	0
			RIS	10.65	11.00	1.84	7.00	15.00	0	0
			GLMERT	3.86	4.00	0.98	1.00	6.00	2	2
			GMERT	3.08	3.00	1.69	1.00	8.00	8	1
			BiMM	3.57	3.00	0.80	3.00	6.00	0	0
7	INTERCEPT & SLOPE Low	Large	Std	6.19	6.00	1.85	4.00	11.00	31	15
			RI	9.73	10.00	1.95	5.00	13.00	0	0
			RIS	9.57	10.00	1.99	5.00	13.00	0	0
			GLMERT	6.27	6.00	0.61	5.00	8.00	80	52
			GMERT	6.30	6.00	0.81	5.00	9.00	70	30
			BiMM	3.14	3.00	0.63	2.00	5.00	0	0
8	High		Std	6.95	6.00	4.56	1.00	18.00	0	0
			RI	9.30	9.00	1.79	5.00	12.00	0	0
			RIS	9.97	10.00	1.48	8.00	13.00	0	0
			GLMERT	4.95	5.00	1.35	3.00	8.00	23	15
			GMERT	4.92	5.00	1.82	2.00	9.00	53	23
			BiMM	3.54	3.00	0.90	2.00	6.00	3	0
9	Low	Small	Std	7.35	6.00	3.17	2.00	15.00	21	5
			RI	10.27	11.00	1.79	7.00	13.00	0	0
			RIS	10.30	10.00	1.71	7.00	14.00	0	0
			GLMERT	4.84	5.00	0.96	3.00	7.00	23	0
			GMERT	4.95	5.00	1.82	1.00	10.00	40	16
			BiMM	3.73	4.00	0.73	2.00	5.00	0	0
10	High		Std	4.86	3.00	3.14	1.00	13.00	3	0
			RI	10.30	10.00	1.70	6.00	13.00	0	0
			RIS	10.41	10.00	2.01	6.00	15.00	2	0
			GLMERT	3.89	4.00	0.81	2.00	5.00	3	0
			GMERT	3.35	3.00	1.55	1.00	9.00	9	6
			BiMM	3.49	3.00	0.84	2.00	6.00	0	0

Appendix B.2. Simulations Based on Data with Linear and Non-Linear Fixed Effects

The DGP presented in the simulation of Section 3 is tree-shaped. To complete this simulation, we investigated two other DGPs, the first based on data with linear fixed effects and the second based on data with non-linear fixed effects. Again, 10 different scenarios were considered, involving small/large fixed effects and models with/without random effects. The same cluster configuration and random components shown in Section 3 were used, and the results are based on 100 runs.

Again, we followed the DGP presented in [10]. The first DGP had linear fixed effects $f(x_{ij})$. The large-effects scenario used $f(x_{ij}) = 1.20x_{1ij} - 0.3x_{2ij} - 0.2x_{3ij}$, and the small-effects scenario had $f(x_{ij}) = -0.6x_{1ij} - 0.15x_{2ij} - 0.10x_{3ij}$. The results are presented in Table A3. As expected, the *GLMM* had the best predictive performance, since it used the true fixed and random effect structures. Nevertheless, *RI*’s and *RIS*’s performances were very similar to that of *GLMM*, and they outperformed all other methods, for all random effects scenarios. This is the best result regarding the GMET method, which when data had a linear structure, thanks to its step involving a GLM model, had outstanding performances.

The second DGP had non-linear fixed effects $f(x_{ij})$. The large-effects scenario used $f(x_{ij}) = 1.0x_{2ij} - 0.60x_{2ij}^2 - 4.80(x_{3ij} > 0) + 0.80x_{1ij}x_{3ij}$; the small-effects scenario had $f(x_{ij}) = 0.50x_{2ij} - 0.30x_{2ij}^2 - 2.40(x_{3ij} > 0) + 0.40x_{1ij}x_{3ij}$. The results are presented in Table A4. Again, *GLMM* had the best predictive performance, followed by *GLMERT*. *GMET* and *GMERT* had similar performances that increased when random effects were high; they got very close to *GLMERT*;s performance.

Table A3. Results of the 100 simulation runs in terms of predictive probability mean absolute deviation (PMAD) and predictive misclassification rate (PMCR) of the seven models for the 10 DGPs based on data with linear fixed effect. *GMET* outperformed all other tree-based methods.

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
1	NO RANDOM EFFECT	Large	Std	10.25	10.53	1.30	8.12	12.29	12.97	12.84	1.21	10.76	15.44
			RI	7.38	7.38	0.60	6.15	8.46	13.55	13.28	2.14	11.16	23.08
			RIS	7.41	7.50	0.61	6.19	8.34	12.49	12.64	0.72	10.92	13.68
			GLMM	3.26	3.32	0.67	2.02	4.55	11.08	11.06	0.80	9.28	12.40
			GLMERT	8.32	8.20	0.73	6.94	9.54	16.54	14.36	5.10	12.32	35.24
			GMERT	11.65	11.46	0.97	9.79	13.58	13.32	13.18	1.16	11.56	16.20
			BiMM	10.88	10.94	0.98	8.61	12.96	13.24	13.48	1.03	10.56	15.00
2		Small	Std	4.97	4.95	0.49	4.12	6.31	4.72	4.64	0.51	3.80	6.36
			RI	2.85	2.75	0.52	1.90	3.91	6.19	5.76	1.51	4.24	9.80
			RIS	2.95	2.89	0.59	2.20	4.23	7.14	6.92	1.28	4.32	9.20
			GLMM	2.51	2.43	0.61	1.40	3.76	7.39	7.10	1.34	5.20	11.40
			GLMERT	3.30	3.15	0.60	2.55	4.66	6.17	6.28	1.40	3.80	9.04
			GMERT	7.60	7.49	0.50	6.68	8.64	6.40	6.12	1.72	4.20	9.80
			BiMM	5.01	4.97	0.47	4.26	6.31	4.65	4.62	0.40	3.80	5.80
3	Low	Large	Std	16.22	16.24	1.46	13.63	19.25	17.15	16.80	1.17	15.28	20.20
			RI	9.50	9.49	0.68	7.92	10.91	13.47	13.36	1.03	11.48	15.44
			RIS	9.35	9.26	0.74	7.83	10.82	13.35	13.20	1.07	11.52	15.44
			GLMM	6.65	6.67	0.73	5.32	7.76	11.92	11.94	1.00	9.76	13.92
			GLMERT	10.31	10.27	0.84	8.86	12.87	14.23	14.32	1.02	12.44	16.20
			GMERT	16.80	16.77	0.61	15.70	18.07	15.78	15.84	1.35	13.24	18.44
			BiMM	16.43	16.21	1.23	14.14	18.59	17.19	16.86	1.14	15.32	19.08
4	High		Std	21.14	21.37	2.44	14.17	25.81	21.28	21.66	1.78	17.64	24.56
			RI	9.89	9.88	0.76	8.26	11.92	13.54	13.46	1.02	11.56	16.52
			RIS	9.67	9.60	0.69	8.09	11.07	13.17	13.20	0.98	11.08	15.24
			GLMM	7.18	7.16	0.69	5.58	8.45	11.50	11.62	1.01	9.28	13.40
			GLMERT	10.76	10.92	0.82	8.96	12.18	14.18	14.18	0.97	12.32	15.84
			GMERT	21.29	21.51	1.73	16.31	24.91	18.85	18.72	2.73	14.92	27.80
			BiMM	21.18	21.37	2.31	14.63	25.81	20.85	20.56	1.81	16.44	24.56
	INTERCEPT												

Table A3. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
5	Low	Small	Std	5.95	5.93	0.62	4.59	7.38	5.30	5.28	0.65	4.24	6.72
			RI	3.69	3.64	0.66	2.63	5.76	7.02	7.02	1.62	4.36	10.08
			RIS	3.73	3.70	0.73	2.94	6.80	7.94	7.88	1.27	5.52	10.60
			GLMM	3.30	3.16	0.68	2.41	6.01	8.13	8.06	1.04	6.40	11.80
			GLMERT	4.07	4.05	0.54	3.10	5.29	6.91	6.98	1.55	4.36	9.20
			GMERT	8.32	8.26	0.49	7.46	9.40	6.91	7.08	1.32	4.36	9.36
			BiMM	5.95	5.93	0.62	4.59	7.38	5.30	5.28	0.65	4.24	6.72
6	High		Std	12.28	12.40	2.00	8.26	15.67	9.98	9.92	1.76	6.32	13.84
			RI	5.80	5.82	0.83	4.01	7.35	9.99	9.92	1.88	6.88	14.24
			RIS	5.78	5.83	0.83	4.00	7.22	9.84	9.84	1.70	6.88	13.20
			GLMM	5.03	4.98	0.78	3.41	6.65	9.49	9.48	1.61	6.80	13.12
			GLMERT	6.40	6.26	0.93	4.78	8.13	10.36	10.20	1.70	7.36	14.64
			GMERT	12.32	12.34	1.30	9.62	14.62	10.96	10.40	2.42	6.52	17.20
			BiMM	12.30	12.40	1.98	8.26	15.67	9.88	9.86	1.77	6.32	13.84
7	Low	Large	Std	14.96	15.08	1.41	11.61	17.73	16.31	16.42	1.23	13.56	18.16
			RI	9.52	9.42	0.70	8.17	11.26	14.05	14.04	0.96	12.08	16.48
			RIS	9.66	9.65	0.71	8.30	11.31	14.04	14.06	0.95	12.56	16.52
			GLMM	6.77	6.69	0.60	5.71	7.98	12.62	12.66	0.89	11.08	14.40
			GLMERT	10.73	10.79	0.93	8.95	12.47	14.90	14.84	1.15	12.72	17.00
			GMERT	15.50	15.62	1.05	13.52	17.77	15.61	15.26	1.17	13.44	18.88
			BiMM	15.20	15.12	1.35	11.91	17.73	16.86	16.90	1.34	13.56	19.92
8	High		Std	23.07	22.76	2.58	19.28	29.66	22.32	22.24	2.57	17.64	28.80
			RI	10.71	10.44	1.11	9.11	13.27	14.47	14.36	1.69	11.76	19.12
			RIS	10.56	10.38	1.15	8.96	14.22	14.63	14.62	1.49	12.20	18.40
			GLMM	7.97	8.02	1.02	6.22	10.38	12.93	12.96	1.33	10.76	16.60
			GLMERT	12.13	11.77	0.97	10.55	14.80	16.00	15.84	1.65	12.92	20.32
			GMERT	18.52	18.57	1.01	16.82	21.02	18.77	18.70	2.12	15.76	23.68
	INTERCEPT & SLOPE		BiMM	23.20	23.08	2.49	19.45	29.66	22.76	22.50	2.62	17.32	28.56
9	Low	Small	Std	5.65	5.57	0.65	4.71	7.35	5.09	5.16	0.52	3.92	6.20
			RI	3.46	3.42	0.54	2.39	4.44	6.87	6.78	1.69	4.48	10.16
			RIS	3.63	3.73	0.61	2.68	4.80	7.71	7.64	0.99	6.28	10.44
			GLMM	3.14	3.08	0.67	2.01	4.67	8.02	7.76	1.13	6.16	10.52
			GLMERT	3.88	3.72	0.67	3.01	5.59	8.23	7.80	1.35	5.20	11.00
			GMERT	8.02	8.03	0.45	7.13	8.98	8.42	8.32	1.23	6.36	11.00
			BiMM	5.68	5.59	0.64	4.79	7.35	5.10	5.20	0.52	3.92	6.20
10	High		Std	9.34	9.31	1.49	6.33	13.26	7.93	7.84	1.49	5.00	10.68
			RI	5.63	5.54	0.93	3.84	7.31	9.69	9.84	1.49	7.00	12.32
			RIS	5.71	5.61	0.95	3.70	7.45	9.94	10.24	1.71	6.36	13.88
			GLMM	5.14	5.19	0.96	3.11	7.72	9.79	9.92	1.56	6.40	12.92
			GLMERT	5.89	5.90	0.83	4.33	7.48	10.09	10.08	1.35	7.16	12.76
			GMERT	10.56	10.50	0.86	9.07	12.66	10.97	11.00	1.99	8.04	15.24
			BiMM	9.34	9.31	1.49	6.33	13.26	7.93	7.84	1.49	5.00	10.68

Table A4. Results of the 100 simulation runs in terms of predictive probability mean absolute deviation (PMAD) and predictive misclassification rate (PMCR) of the seven models for the 10 DGPs based on data with non-linear fixed effect.

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
1	NO RANDOM EFFECT	Large	Std	14.96	14.96	1.71	11.15	17.79	12.17	12.60	1.25	9.32	14.08
			RI	17.75	17.90	1.65	13.59	20.18	15.63	15.56	1.74	12.08	20.32
			RIS	18.00	17.91	1.80	14.00	21.95	15.67	15.52	1.69	11.96	20.72
			GLMM	9.45	9.49	0.72	7.91	10.55	8.49	8.44	0.67	7.36	9.72
			GLMERT	12.34	12.34	1.35	9.75	15.06	11.90	12.00	1.34	9.24	14.88
			GMERT	17.59	17.35	1.12	15.77	21.03	12.53	12.52	1.31	9.76	16.00
			BiMM	26.12	25.60	2.90	21.92	31.53	47.78	47.74	0.87	45.96	49.68
2		Small	Std	14.62	14.38	1.67	11.81	17.83	13.16	13.00	1.37	11.12	16.44
			RI	16.74	16.72	1.58	13.80	20.19	16.89	16.68	1.83	14.28	20.88
			RIS	16.54	16.67	1.31	12.79	18.84	16.18	16.16	1.44	14.16	19.76
			GLMM	9.14	9.22	0.58	7.88	10.46	9.77	10.00	0.56	8.48	10.64
			GLMERT	13.54	13.52	1.23	11.17	16.03	14.24	13.80	2.01	11.72	19.24
			GMERT	16.33	15.94	1.11	14.58	18.89	13.51	13.28	1.55	11.28	18.00
			BiMM	24.75	23.95	2.91	19.99	31.41	48.21	48.46	1.05	45.80	50.00
3	INTERCEPT Low	Large	Std	18.03	17.71	2.29	14.29	26.00	14.94	14.56	2.09	11.24	20.68
			RI	18.01	17.90	1.95	14.88	21.99	15.61	15.40	1.77	12.32	18.76
			RIS	17.71	17.88	1.79	14.25	21.09	15.63	15.84	1.66	12.88	18.80
			GLMM	10.20	9.90	0.86	8.95	11.94	8.96	8.92	0.77	7.52	10.12
			GLMERT	13.62	13.33	1.34	11.45	16.56	12.98	13.16	1.43	10.88	16.20
			GMERT	18.25	18.22	0.96	16.13	20.06	13.25	12.92	1.31	10.36	16.44
			BiMM	26.38	25.88	3.25	21.34	33.48	47.99	48.06	1.18	44.64	49.96
4	High		Std	18.60	18.82	1.71	15.48	20.16	15.48	15.72	1.32	13.24	16.88
			RI	17.94	17.40	1.84	14.74	21.87	15.44	15.60	1.24	13.28	17.36
			RIS	17.67	17.21	1.71	14.56	20.34	15.61	15.68	1.27	12.96	18.08
			GLMM	10.53	10.59	0.86	8.96	12.61	9.31	9.44	0.71	7.92	11.12
			GLMERT	14.45	14.51	1.29	11.98	16.84	13.86	13.88	0.94	12.16	16.12
			GMERT	18.97	18.97	0.98	17.07	21.15	13.84	13.64	1.38	11.16	15.92
			BiMM	27.76	26.56	3.15	23.03	34.41	48.01	48.18	1.13	45.12	50.04
5	Low	Small	Std	15.36	15.64	1.91	12.43	20.27	14.03	14.00	1.65	10.72	18.92
			RI	17.08	17.11	1.70	14.25	21.35	16.83	16.32	1.80	13.96	20.04
			RIS	16.41	16.39	1.56	14.22	20.63	16.23	16.00	1.75	13.12	20.16
			GLMM	9.46	9.28	0.81	7.44	11.50	10.03	10.04	0.86	8.12	11.40
			GLMERT	13.35	13.13	1.18	11.56	16.37	14.30	14.20	1.60	11.04	17.96
			GMERT	17.05	16.96	0.96	14.97	19.62	14.45	14.48	1.23	10.80	17.08
			BiMM	25.46	25.29	2.75	18.59	29.47	48.29	48.50	1.18	44.96	50.56
6	High		Std	17.50	17.64	1.99	12.66	21.25	15.73	15.76	1.60	12.40	18.56
			RI	16.77	16.82	1.14	14.13	18.56	16.39	16.52	0.96	14.28	18.16
			RIS	16.92	16.78	1.63	12.96	21.42	16.47	16.36	1.27	13.92	18.48
			GLMM	10.46	10.38	0.62	9.26	12.18	11.05	10.88	0.71	10.04	12.52
			GLMERT	14.52	14.62	1.06	12.05	16.77	15.36	15.44	1.45	12.48	18.28
			GMERT	18.77	18.64	1.13	16.46	21.04	15.93	15.96	1.51	12.84	18.56
			BiMM	26.89	26.62	2.90	21.34	31.80	48.45	48.52	1.38	44.60	51.60

Table A4. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)					PMCR (%)				
				Mean	Median	SD	Min	Max	Mean	Median	SD	Min	Max
7	INTERCEPT & SLOPE Low	Large	Std	15.87	15.25	2.38	12.04	22.45	12.72	12.52	1.70	9.80	17.68
			RI	17.58	17.65	1.51	14.46	19.90	15.40	15.40	1.30	12.68	18.12
			RIS	17.36	17.55	1.39	14.48	19.54	15.48	15.48	1.31	12.68	18.56
			GLMM	9.97	9.86	0.84	8.49	11.52	9.21	8.92	0.84	7.76	10.92
			GLMERT	13.48	13.29	1.26	11.02	16.17	13.18	13.16	1.63	9.80	16.56
			GMERT	18.27	18.33	0.97	16.55	20.16	13.03	13.00	1.51	10.28	15.32
			BiMM	25.98	25.76	2.70	20.66	30.75	47.92	48.04	1.09	44.00	49.16
8	High		Std	16.98	16.77	2.03	13.43	20.19	13.76	14.00	1.49	10.92	17.72
			RI	18.39	17.82	1.84	14.84	21.40	15.85	16.08	1.66	11.60	18.92
			RIS	17.84	17.70	1.81	14.51	20.98	15.69	15.80	1.53	12.08	18.84
			GLMM	10.48	10.47	0.92	8.55	12.19	9.73	9.60	0.84	8.36	11.52
			GLMERT	14.27	14.16	1.04	12.40	16.59	13.51	13.32	1.41	11.20	17.20
			GMERT	19.22	19.19	1.35	16.74	22.30	13.99	13.92	1.74	11.24	17.80
			BiMM	27.10	26.19	3.36	21.90	34.64	47.65	47.64	1.27	44.64	50.44
9	Low	Small	Std	15.24	14.99	1.67	12.15	19.15	13.76	13.40	1.09	11.96	16.04
			RI	16.54	16.36	1.63	13.90	21.18	16.01	15.92	1.10	13.52	18.72
			RIS	16.57	16.44	1.51	13.88	20.16	16.13	16.24	1.11	13.52	18.80
			GLMM	9.47	9.57	0.54	8.13	10.37	10.16	10.16	0.69	8.84	11.72
			GLMERT	13.33	13.00	1.08	11.45	15.47	14.35	14.28	1.36	10.76	16.76
			GMERT	16.84	16.81	1.07	15.41	19.88	14.22	14.40	1.18	12.48	17.24
			BiMM	25.95	25.96	3.40	20.12	32.32	48.01	48.10	0.94	45.16	49.64
10	High		Std	17.04	17.00	1.96	13.83	23.24	15.44	15.44	1.46	12.92	19.28
			RI	16.18	16.41	1.56	13.81	19.33	15.32	16.96	1.31	13.84	17.68
			RIS	16.12	16.28	1.08	14.35	17.91	15.18	15.32	1.36	12.88	18.44
			GLMM	10.44	10.54	0.61	9.19	11.31	11.06	11.08	0.72	9.36	12.36
			GLMERT	14.59	14.69	1.27	12.38	18.56	15.32	15.16	1.50	13.12	19.76
			GMERT	18.61	18.40	1.18	15.87	21.72	15.58	15.24	1.70	11.96	20.16
			BiMM	26.69	26.31	3.01	21.28	34.81	48.66	48.58	1.19	46.00	51.40

Appendix B.3. Simulation Based on Data with a Poisson Response Variable and Unbalanced Clusters

In all the simulations presented in previous sections, we always considered the case of a binary response variable and balanced clusters. Here, to extend the simulation to a broader scenario, we consider DGPs for data with a different response variable in the exponential family, i.e., a Poisson response variable, and unbalanced clusters. We investigated 10 different scenarios involving small/large *linear* fixed-effects and 10 different scenarios involving small/large *tree-shaped* fixed-effects, both cases involving models with/without random effects. Random and fixed components for the 10 DGPs with tree-shaped fixed effects are shown in Table A5. Random components for the DGPs with linear fixed effects were those in Table A5, and linear fixed effects were $f(x_{ij}) = 0.6x_{1ij} + 0.3x_{2ij} + 0.2x_{3ij}$ for the large fixed-effects scenario and $f(x_{ij}) = 0.3x_{1ij} + 0.15x_{2ij} + 0.1x_{3ij}$ for the small fixed-effects scenario. The variables $X_1, \ldots, X_8$ were generated as uniformly distributed on the interval $[0, 5]$. Regarding the cluster configuration, we simulated a new scenario in which the 50 clusters were unbalanced, while considering 10 different cluster sizes. In particular, the number of observations within clusters took values in $\{62, 64, 66, \ldots, 80\}$, considering five clusters for each pair between 62 and 80. Within each cluster, about 30% of observations were used as the training set and 70% as the test set. For the Poisson response variable simulations, we compared *GMET* with the standard tree (*Std*), *GLMM* and *GLMERT*. We omitted *BiMM* because it can handle only a binary response variable and *GMERT* because the code did not work for a Poisson family distribution. Results in terms of PMAD are reported in Tables A6 and A7 for the tree-shaped and linear fixed effects, respectively.

By looking at Table A6, we can observe that *GLMERT* still had the best performances. Nonetheless, *GMET*'s performances were very close to those of *GLMM* in DGPs 1–6. *GMET* performed better than *GLMM* in DGPs 7–10. Lastly, by according to Table A7, *GMET* had good performance compared to other tree-based methods, when data had a linear structure. Indeed, except for DGPs 1 and 7, *GMET* was always second to *GLMM*, outperforming *GLMERT* and *Std*.

Table A5. Data generating processes (DGP) for the simulation study with a Poisson response variable.

DGP	Random Component				Fixed Component						
	Structure	Effect	ψ_{11}	ψ_{22}	Effect	φ^1	φ^2	φ^3	φ^4	φ^5	φ^6
1	No random effect	–	–	–	Large	4	6	8	6	4	10
2		–			Small	2	4	6	4	2	8
3	Random Intercept	Low	2.00	–	Large	4	6	8	6	4	10
4		High	5.00	–							
5		Low	0.25	–	Small	2	4	6	4	2	8
6		High	2.00	–							
7	Random Intercept and Slope	Low	2.00	0.05	Large	4	6	8	6	4	10
8		High	5.00	0.25							
9		Low	0.25	0.01	Small	2	4	6	4	2	8
10		High	2.00	0.05							

Table A6. Results of the 100 simulation runs in terms of predictive probability mean absolute deviation (PMAD) of the five models for the 10 DGPs with a Poisson response variable and tree-shaped fixed effects.

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)				
				Mean	Median	SD	Min	Max
1	NO RANDOM EFFECT	Large	Std	2.89	2.85	1.75	0.07	6.10
			RI	10.58	10.49	2.34	5.33	16.91 4
			RIS	11.15	11.18	2.34	5.36	16.91
			GLMM	8.46	8.44	2.28	4.44	16.59
			GLMERT	3.99	3.22	2.66	0.30	11.04
2		Small	Std	4.64	2.82	3.95	1.31	17.62
			RI	16.57	16.33	3.76	8.59	28.19
			RIS	16.76	16.25	3.92	8.71	28.65
			GLMM	12.89	12.16	3.19	7.79	22.97
			GLMERT	5.96	4.75	4.55	1.31	17.63
3	Low	Large	Std	557.78	551.49	228.98	229.48	1351.42
			RI	32.26	32.79	5.70	20.91	44.37
			RIS	32.86	33.33	5.77	21.13	45.65
			GLMM	30.95	30.69	6.37	20.57	47.55
			GLMERT	27.12	26.38	5.82	18.10	42.41
4	High INTERCEPT		Std	2920.44	2265.68	2382.02	412.82	12318.30
			RI	52.60	52.36	13.56	21.34	86.19
			RIS	54.28	55.26	13.95	21.65	91.11
			GLMM	49.08	47.20	13.07	21.94	84.06
			GLMERT	38.95	36.60	12.34	20.89	71.20
5	Low	Small	Std	176.50	175.42	29.01	131.43	271.54
			RI	32.16	32.50	3.37	23.73	38.54
			RIS	32.30	32.48	3.30	24.15	38.47
			GLMM	31.27	31.86	3.56	22.82	36.21
			GLMERT	28.59	29.27	3.77	19.81	34.86
6	High		Std	1074.88	982.37	394.96	519.43	2068.50
			RI	42.26	42.55	5.45	31.33	57.30
			RIS	45.62	45.47	6.19	33.86	59.50
			GLMM	41.74	41.60	5.72	29.47	54.00
			GLMERT	36.77	37.38	5.70	26.49	52.03

Table A6. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)				
				Mean	Median	SD	Min	Max
7	Low	Large	Std	768.61	661.59	395.50	256.47	1842.23
			RI	148.34	132.52	59.64	63.67	330.14
			RIS	42.36	41.44	8.80	26.37	68.05
			GLMM	41.80	43.44	6.90	28.98	56.31
			GLMERT	38.71	39.91	7.87	25.13	58.30
8	High INTERCEPT & SLOPE		Std	8197.50	5322.26	9631.35	1428.45	47,610.80
			RI	2010.04	1170.11	2150.09	250.94	10,877.54
			RIS	85.43	83.32	27.42	39.57	158.14
			GLMM	89.12	77.31	41.90	38.69	265.94
			GLMERT	71.17	65.29	22.84	43.63	138.36
9	Low	Small	Std	206.85	200.50	32.50	151.81	279.47
			RI	61.09	60.30	8.71	44.70	81.25
			RIS	41.40	41.24	4.17	34.27	51.55
			GLMM	41.23	40.29	3.63	36.25	50.12
			GLMERT	38.79	38.25	3.76	33.00	45.77
10	High		Std	1570.61	1331.61	1068.54	503.47	5989.87
			RI	303.01	247.81	166.30	147.81	893.52
			RIS	61.23	61.58	14.25	40.12	103.83
			GLMM	62.74	62.20	14.53	40.53	113.51
			GLMERT	56.69	55.13	12.32	35.43	93.44

Table A7. Results of the 100 simulation runs in terms of predictive probability mean absolute deviation (PMAD) of the five models for the 10 DGPs with a Poisson response variable and linear fixed effects.

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)				
				Mean	Median	SD	Min	Max
1	NO RANDOM EFFECT	Large	Std	167.20	165.86	6.16	158.25	180.22
			RI	161.46	161.06	5.15	147.96	171.24
			RIS	154.84	157.35	9.30	137.29	167.30
			GLMM	19.37	19.39	4.00	11.63	29.76
			GLMERT	137.01	137.46	5.85	124.88	147.33
2		Small	Std	35.52	34.66	3.52	29.49	46.30
			RI	26.52	26.18	1.38	24.35	29.55
			RIS	26.84	26.78	1.61	24.37	30.52
			GLMM	11.31	11.23	2.07	6.83	15.96
			GLMERT	34.38	34.48	2.86	29.50	39.94
3	Low	Large	Std	4721.44	4580.55	1647.63	2531.04	10,154.36
			RI	955.46	884.97	294.72	559.66	1895.78
			RIS	812.60	751.71	236.17	489.89	1600.11
			GLMM	86.95	84.49	13.71	64.38	119.16
			GLMERT	748.79	725.66	237.72	395.91	1536.27
4	High INTERCEPT		Std	29,879.11	23,361.45	28,903.16	8831.63	163,613.27
			RI	4765.29	3490.10	4220.57	1488.83	23,464.90
			RIS	4020.26	3191.48	3262.35	1404.12	17,930.40
			GLMM	161.18	152.26	68.80	93.19	497.72
			GLMERT	4463.56	3063.96	4394.33	1184.40	22,811.63

Table A7. *Cont.*

DGP	Random Effect	Fixed Effect	Fitted Model	PMAD (%)				
				Mean	Median	SD	Min	Max
5	INTERCEPT & SLOPE Low	Small	Std	196.94	187.97	28.00	146.76	249.27
			RI	58.32	57.60	4.36	51.28	68.91
			RIS	59.01	58.07	4.46	51.45	68.53
			GLMM	32.59	32.56	3.14	27.64	39.31
			GLMERT	63.72	63.18	4.26	57.68	70.82
6	High		Std	1076.42	898.49	617.26	518.63	3709.94
			RI	135.98	117.49	99.36	76.91	636.51
			RIS	116.19	113.54	32.04	74.79	195.33
			GLMM	41.82	39.14	8.56	28.96	67.62
			GLMERT	122.63	104.09	56.03	74.17	376.44
7	Low	Large	Std	7208.83	6637.47	3058.53	2687.22	15,020.33
			RI	1588.80	1449.69	583.86	710.16	2803.83
			RIS	1181.85	1085.58	432.46	524.84	2208.55
			GLMM	133.09	130.27	22.26	96.65	187.50
			GLMERT	496.32	442.54	193.57	263.06	1049.15
8	High		Std	114,321.97	66,083.74	177,762.92	8903.18	909,771.14
			RI	28,423.89	15,363.83	43,872.04	2250.90	207,622.27
			RIS	11,753.23	5637.47	20,408.21	1366.25	107,551.55
			GLMM	305.21	255.73	145.78	127.61	811.44
			GLMERT	11,844.52	4530.65	22,981.80	561.41	10,7891.01
9	Low	Small	Std	236.25	230.69	42.66	161.23	361.97
			RI	70.23	70.95	8.95	55.76	86.57
			RIS	63.67	63.89	6.76	51.49	77.41
			GLMM	38.91	39.50	4.25	29.06	46.83
			GLMERT	69.70	69.83	6.32	57.10	82.35
10	High		Std	2069.59	1778.77	1740.42	437.63	9234.87
			RI	312.70	257.02	218.08	96.06	1201.62
			RIS	112.36	122.74	99.71	71.13	427.66
			GLMM	63.24	61.68	12.61	45.85	92.99
			GLMERT	125.77	108.76	77.33	63.03	464.06

Notes

[1] Universitat Autonoma de Barcelona (UAB)—Spain; Instituto Politecnico de Braganca (IPB)—Portugal; Opole University of Technology—Poland; Politecnico di Milano—Italy; Universidad de Leon—Spain; University of Galati *Dunarea de Jos*—Romania.

[2] In particular, the proposed method can deal with response variables that belong to the following families: binomial, Gaussian, gamma, inverse-Gaussian, Poisson, quasi, quasi-binomial, quasi-Poisson (i.e., the distributions handled by GLMM).

[3] Fixed-effects covariates, random effect coefficents and binary response variables were generated using the `runif()`, `rnorm()` and `rbinom()` functions implemented R software, respectively. Parameters of these functions are reported in Figure 1 and Table 1.

[4] The random intercept was the only random effect structure that BiMM algorithm handled.

[5] We chose 20 as the minimum number of observations to attempt a split because it is the default number within the *rpart* R package; 10 as maximum depth was chosen in order not to grow "overly large" trees, but interpretable ones. The final depth of each tree was chosen by cross-validation (the complexity parameter of the tree was automatically chosen by cross-validation within the algorithm), and it was always smaller than 10.

[6] This might have also been due to the fact that *BiMM* was disadvantaged, since it does not handle a random slope but only a random intercept.

References

1. SPEETproject. SPEET, Proposal for Strategic Partnerships (Proposal Narrative). 2017. Available online: https://www.speet-project.com/the-project (accessed on 5 May 2020).
2. Barbu, M.; Vilanova, R.; Lopez Vicario, J.; Pereira, M.J.; Alves, P.; Podpdora, M.; Ángel Prada, M.; Morán, A.; Torreburno, A.; Marin, S.; et al. Data mining tool for academic data exploitation: Literature review and first architecture proposal. In *Projecto SPEET-Student Profile for Enhancing Engineering Tutoring*; IEEE Access: Piscataway, NJ, USA, 2017.
3. Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2010**, *40*, 601–618. [CrossRef]
4. Bock, R.D. *Multilevel Analysis of Educational Data*; Elsevier: London, UK, 2014.
5. Goldstein, H. *Multilevel Statistical Models*; John Wiley & Sons: lWest Sussex, UK, 2011; Volume 922.
6. Agresti, A. *An Introduction to Categorical Data Analysis*; Wiley: Hoboken, NJ, USA, 2018.
7. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees, The Wadsworth Statistics and Probability Series*; Wadsworth International Group: Belmont, CA, USA, 1984; p. 356.
8. Sela, R.J.; Simonoff, J.S. RE-EM trees: A data mining approach for longitudinal and clustered data. *Mach. Learn.* **2012**, *86*, 169–207. [CrossRef]
9. Hajjem, A.; Bellavance, F.; Larocque, D. Mixed effects regression trees for clustered data. *Stat. Probab. Lett.* **2011**, *81*, 451–459. [CrossRef]
10. Hajjem, A.; Larocque, D.; Bellavance, F. Generalized mixed effects regression trees. *Stat. Probab. Lett.* **2017**, *126*, 114–118. [CrossRef]
11. Fokkema, M.; Smits, N.; Zeileis, A.; Hothorn, T.; Kelderman, H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav. Res. Methods* **2018**, *50*, 2016–2034. [CrossRef]
12. Speiser, J.L.; Wolf, B.J.; Chung, D.; Karvellas, C.J.; Koch, D.G.; Durkalski, V.L. BiMM tree: A decision tree method for modeling clustered and longitudinal binary outcomes. In *Communications in Statistics-Simulation and Computation*; Taylor & Francis: Boca Raton, FL, USA, 2020 ; Volume 49, pp. 1–20.
13. Zeileis, A.; Hothorn, T.; Hornik, K. Model-based recursive partitioning. *J. Comput. Graph. Stat.* **2008**, *17*, 492–514. [CrossRef]
14. Cabrera, A.F.; Stampen, J.O.; Hansen, W.L. Exploring the effects of ability to pay on persistence in college. *Rev. High. Educ.* **1990**, *13*, 303–336. [CrossRef]
15. John, E.P.S.; Paulsen, M.B.; Starkey, J.B. The nexus between college choice and persistence. *Res. High. Educ.* **1996**, *37*, 175–220. [CrossRef]
16. Pascarella, E.T.; Terenzini, P.T. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *J. High. Educ.* **1980**, *51*, 60–75. [CrossRef]
17. Spady, W.G. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange* **1970**, *1*, 64–85. [CrossRef]
18. Tinto, V. Dropout from higher education: A theoretical synthesis of recent research. *Rev. Educ. Res.* **1975**, *45*, 89–125. [CrossRef]
19. Korhonen, V.; Rautopuro, J. Identifying problematic study progression and "at-risk" students in higher education in Finland. *Scand. J. Educ. Res.* **2019**, *63*, 1056–1069. [CrossRef]
20. Seidel, E.; Kutieleh, S. Using predictive analytics to target and improve first year student attrition. *Aust. J. Educ.* **2017**, *61*, 200–218. [CrossRef]
21. Sothan, S. The determinants of academic performance: Evidence from a Cambodian university. *Stud. High. Educ.* **2019**, *44*, 2096–2111. [CrossRef]
22. Saa, A.A.; Al-Emran, M.; Shaalan, K. Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technol. Knowl. Learn.* **2019**, *24*, 567–598.
23. Wook, M.; Yusof, Z.M.; Zakree, M.; Nazri, A. Educational data mining acceptance among undergraduate students. *Educ. Inf. Technol.* **2017**, *22*, 1195. [CrossRef]
24. Tampakas, V.; Livieris, I.E.; Pintelas, E.; Karacapilidis, N.; Pintelas, P. Prediction of students' graduation time using a two-level classification algorithm. In Proceedings of the International Conference on Technology and Innovation in Learning, Teaching and Education, Thessaloniki, Greece, 20–22 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 553–565.
25. Sanyal, D.; Bosch, N.; Paquette, L. Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models. In *International Educational Data Mining Society*; ERIC, 2020.
26. Sivakumar, S.; Venkataraman, S.; Selvaraj, R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. *Indian J. Sci. Technol.* **2016**, *9*, 1–5. [CrossRef]
27. Yasmin, D. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Educ.* **2013**, *34*, 218–231. [CrossRef]
28. Abu-Oda, G.S.; El-Halees, A.M. Data mining in higher education: University student dropout case study. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*. [CrossRef]
29. Meedech, P.; Iam-On, N.; Boongoen, T. Prediction of student dropout using personal profile and data mining approach. In *Intelligent and Evolutionary Systems*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 143–155.
30. Team, R.C. *R: A Language and Environment for Statistical Computing;* R Foundation for Statistical Computing: Vienna, Austria, 2014.
31. Searle, S.R.; McCulloch, C.E. *Generalized, Linear, and Mixed Models*; Wiley: Hoboken, NJ, USA, 2001.
32. McCullagh, P.; Nelder, J. *Generalized Linear Models*; Taylor & Francis Group: Boca Raton, FL, USA, 2019.

33. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2001; Volume 1.
34. Therneau, T.; Atkinson, B.; Ripley, B. Rpart: Recursive Partitioning and Regression Trees (R Package), 2015. Available online: cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf (accessed on 20 April 2016).
35. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *arXiv* **2014**, arXiv:1406.5823.
36. Gueorguieva, R. A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Stat. Model.* **2001**, *1*, 177–193. [CrossRef]
37. Handayani, D.; Notodiputro, K.A.; Sadik, K.; Kurnia, A. A comparative study of approximation methods for maximum likelihood estimation in generalized linear mixed models (GLMM). In Proceedings of the AIP Conference, Jawa Barat, Indonesia, 27–28 September 2016; AIP Publishing LLC: Melville, NY, USA, 2017 ; Volume 1827, p. 020033.
38. Pinheiro, J.; Bates, D. *Mixed-Effects Models in S and S-PLUS*; Springer Science & Business Media: New York, USA, 2006.
39. Goldstein, H.; Browne, W.; Rasbash, J. Partitioning variation in multilevel models. *Underst. Stat. Stat. Issues Psychol. Educ. Soc. Sci.* **2002**, *1*, 223–231. [CrossRef]
40. Browne, W.J.; Subramanian, S.V.; Jones, K.; Goldstein, H. Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2005**, *168*, 599–613. [CrossRef]
41. Pintelas, E.; Livieris, I.E.; Pintelas, P. A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. *Algorithms* **2020**, *13*, 17. [CrossRef]