

## DICHIARAZIONI SOSTITUTIVE DI ATTO DI NOTORIETÀ

AI SENSI DELL'ART. 47 DEL D.P.R. 28 DICEMBRE 2000, N. 445.

La sottoscritta CHIARA MASCI, codice fiscale MSCCHR90S51F205I, nata a Milano (MI) il 11 novembre 1990, di sesso femminile, residente in Via Pasubio, 73, C.A.P. 20081, Abbiategrasso (MI), mail [chiara.masci9@gmail.com](mailto:chiara.masci9@gmail.com), consapevole delle sanzioni penali richiamate dall'art. 76 del D.P.R. 28.12.2000, n. 445 per le ipotesi di falsità in atti e dichiarazioni mendaci

### DICHIARA

Che la pubblicazione

Pellagatti, M., Masci, C., Ieva, F., & Paganoni, A. M. (2021). Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14(3), 241-257, ISSN: 1932-1864, doi: 10.1002/sam.11505

è frutto della collaborazione degli autori che ne condividono contributi e responsabilità. Si dichiara, nello specifico, che Chiara Masci ha contribuito in modo sostanziale alla formulazione dell'idea originale, allo studio di simulazione nella Sezione 3 e ha curato la supervisione del lavoro e parzialmente la stesura del testo. Massimo Pellagatti ha contribuito in modo sostanziale agli sviluppi metodologici e all'implementazione del modello, alle analisi statistiche e alla stesura del testo. Francesca Ieva e Anna Maria Paganoni hanno partecipato a tutte le fasi curando la supervisione del lavoro.

## RESEARCH ARTICLE

# Generalized mixed-effects random forest: A flexible approach to predict university student dropout

Massimo Pellagatti | Chiara Masci<sup>ORCID</sup> | Francesca Ieva<sup>ORCID</sup> | Anna M. Paganoni<sup>ORCID</sup>MOX—Department of Mathematics,  
Politecnico di Milano, Milan, Italy**Correspondence**Anna M. Paganoni, MOX—Department of  
Mathematics, Politecnico di Milano,  
Milan, Italy.  
Email: anna.paganoni@polimi.it**Abstract**

We propose a new statistical method, called generalized mixed-effects random forest (GMERF), that extends the use of random forest to the analysis of hierarchical data, for any type of response variable in the exponential family. The method maintains the flexibility and the ability of modeling complex patterns within the data, typical of tree-based ensemble methods, and it can handle both continuous and discrete covariates. At the same time, GMERF takes into account the nested structure of hierarchical data, modeling the dependence structure that exists at the highest level of the hierarchy and allowing statistical inference on this structure. In the case study, we apply GMERF to Higher Education data to analyze the university student dropout phenomenon. We predict engineering student dropout probability by means of student-level information and considering the degree program students are enrolled in as grouping factor.

**KEYWORDS**

generalized models, hierarchical data, random forest, university students dropout

## 1 | INTRODUCTION

In today's *Big data* era, researchers often have to handle big amounts of complex data. The focus of the analyst is twofold: to reach a high accuracy in the prediction of a studied phenomenon and to understand the complexity of the underlying data structure. The analyst has often to find a compromise between the interpretability of the model, usually high in a simple model, and its accuracy, which often increases as the model complexity increases.

To this purpose, tree-based methods, used for regression and classification, were introduced by Breiman et al.

in [10] and they are now raising in popularity for their high level of interpretability. However, their high variability is often an issue, resulting in poor predictions [21]. New methods using trees as building blocks, called tree-based ensemble methods, started being developed to improve the predictive performance of trees [23]. Bagging, random forest (RF), and boosting are examples of such methods [21]. RF, described in [9], consists in a bootstrap aggregation method that combines the predictions of a large number of trees. In recent years, part of the statistical literature focuses on extending the use of tree-based methods to the analysis of nested data, that is, data with a hierarchical structure, embedding them into mixed-effects models [32]. However, the development of such methods is still at its beginning. One way in which tree-based methods have been extended for modeling nested data is integrating them with linear mixed-effects models (LMMs) [32], with the aim of solving LMMs low-flexibility issue, due to the parametric assumptions.

**Abbreviations:** GLMM, generalized linear mixed model; GMERF, generalized mixed-effects random forest; GMERT, generalized mixed-effects regression tree; LMM, linear mixed-effects model; MERF, mixed-effects random forest; MERT, mixed-effects regression tree; RF, random forest.

LMMs are used to model multilevel data, that is, data in which statistical units naturally have a hierarchical structure (e.g., students nested within schools or patients nested within hospitals), or longitudinal data (e.g., repeated measurements for the same subject). The hierarchical structure of data is worth to be taken into account for several reasons: (a) nested data are not i.i.d., as classical regression or classification models assume, but their distribution depends on their grouping structure; (b) neglecting the hierarchical structure could result in a loss of a valuable piece of data information; (c) disentangling the effects given to each level of the hierarchy allows to understand and to investigate the latent structure present at the highest level of the hierarchy, enriching the knowledge on the phenomenon described by data.

The first method proposed within this context is called mixed-effects regression tree (MERT) [18] and it substitutes the linear combination of the covariates in the fixed effects part of a LMM with a regression tree, built with the same set of covariates. In [35], the authors present an analogous method, but with a different estimation procedure, called random effects expectation maximization tree, that deals with both multilevel and longitudinal data. With the aim of improving the accuracy in predictions, regression trees are replaced by a RF in [19], where the authors develop a method called mixed-effects random forest (MERF). All these methods deal with a Gaussian response variable and they are not suitable to classification problems. Nonetheless, some developments for different types of response variables have also been done. In [20], the MERT approach is extended to non-Gaussian data and a generalized mixed-effects regression tree is proposed. This algorithm is basically the penalized quasi likelihood (PQL) algorithm used to fit generalized linear mixed-effects models (GLMMs) where the weighted linear mixed-effect pseudo-model is replaced by a weighted MERT pseudo-model. Another extension to a classification problem is the generalized mixed-effects tree (GMET), presented in [15], which is in line with the approach of [35] as it uses the tree leaves as indicator variables, rather than using the tree predictions as the MERT approach does. In [14], the authors propose a generalized linear mixed-effects model tree (GLMER tree) algorithm, that alternates the estimates of a GLM tree and a mixed-effects model until convergence. Lastly, the most recent work is proposed in [36], where the authors develop a decision tree method for modeling clustered and longitudinal binary outcomes using a Bayesian setting.

In the context of a non-Gaussian response variable, the existent methods extend the use of simple trees for modeling nested data, but not their ensembles. As we previously state, tree-based methods suffer from high variance and they usually do not have the same level of predictive

**TABLE 1** Tree-based mixed-effects models in the literature

Mixed-effects models	Regression	Classification
Simple tree	MERT [20]	GMERT [20]
	RE-EM trees [35]	GMET [15]
		GLMER tree [14]
Random forest	MERF [19]	GMERF

accuracy as some of the other regression and classification approaches. By aggregating many decision trees, using an ensemble method, the predictive performance of trees can be substantially improved. In this work, we develop a novel model called generalized mixed-effects random forest (GMERF), that is inspired by the GMET model presented in [15], but considers a RF instead of a standard tree in the fixed effects part of the mixed-effects model. This work can then be considered as a further step in the literature about tree-based mixed-effects models as Table 1 illustrates. Following the GMET approach, GMERF is based on a GLMM in which the estimation of the fixed effects part is performed with a RF, with the aim of handling interactions among the different covariates and dealing with highly nonlinear effects. This new method is the first one in the literature able to model hierarchical data with a RF, for a non-Gaussian response variable. Indeed GMERF, as all GLMs, is able to deal with different types of responses, as long as their distribution belongs to the exponential family; this is not true for the Bayesian approach of [36], which works only with binary responses. The strength of this method is that it satisfies the flexibility and the predictive power typical of RF, maintaining the ability of modeling hierarchical data, for different types of response variables in the exponential family. In the recent literature, RF has been extended for new and various statistical tasks: non-parametric quantile regression, conditional average partial effect estimation, and heterogeneous treatment effect estimation via instrumental variables [3]; causal inference [11, 29]; censored quantile regression [24]. To the best of our knowledge, this is the first time that RF is extended to deal with hierarchical data, both for regression and classification, that is, for any response variable in the exponential family.

We describe the GMERF method, providing the pseudo-code of the algorithm for the estimation procedure, we then show a simulation study, comparing its performance to other existing methods and, lastly, we apply it to a case study. We apply GMERF to a real dataset, that Politecnico di Milano selected for the Student Profile of Enhancing Tutoring Engineering (SPEET) project (<https://www.speet-project.com/>). SPEET is a project aimed at determining and categorizing different profiles of engineering students across Europe. SPEET

consortium is composed by six European universities: Universitat Autònoma de Barcelona (UAB)—Barcelona, Spain; Instituto Politecnico de Braganca (IPB)—Braganca, Portugal; Opole University of Technology—Opole, Poland; Politecnico di Milano (PoliMi)—Milano, Italy; Universidad de Leòn—Leòn, Spain; University of Galati *Dunarea de Jos*—Galati, Romania. The essence of SPEET project is to apply data mining algorithms in order to extract information about students and to profile them. A student profile is a set of categories to which a student belongs that gives an insight about how the student is approaching and dealing with his/her studies. Some examples of student profiles are: students that finish degree on time or students that are blocked on a certain set of subjects. Comparisons among different partner institutions will be done in order to establish correlations and get a more complete European-level picture. The role of Politecnico di Milano in the SPEET project is to describe why students leave their studies at the university before accomplishing the degree and to produce a classification method that automatically identifies such students who are likely to drop their studies; from now on we refer to this abandonment as *dropout*. The importance of this task is motivated by the fact that, across all SPEET partners, almost a student out of two leaves his/her engineering studies before obtaining the BSc degree.

In the last decades, the analysis of university students dropout is receiving particular attention in the educational context. Many studies focus on predicting which are the students at risk in the perspective of identifying the determinants of the dropout and of helping those students (see among the others [16, 12, 4, 34]). If it was possible to know as soon as possible to which profile a student belongs, it would be of valuable help for tutors to improve their guiding actions.

We apply GMERF method to Politecnico di Milano data for predicting students dropout probability by means of student-level characteristics and considering the grouping structure of students within engineering degree programs. We consider students as statistical units, grouped based on the degree program they are enrolled in. As student-level covariates, we consider both students performance at Politecnico di Milano (during the first semester of the first year, in the perspective of providing an early warning system) and students collateral data, such as gender, nationality, and previous studies. Results reveal that the dropout is mainly associated to the early performance of the student at university rather than to other student-level variables. Also, with the information at our disposal, we are able to predict the dropout in the 90% of cases.

GMERF represents a breakthrough in the literature of both mixed-effects models and tree-based methods,

combining these two statistical approaches in a robust, flexible, and structured method.

The paper is organized as follows: in Section 2 we describe the GMERF method; in Section 3 we perform a simulation study to investigate the strengths and weaknesses of our method, compared to other existing ones; Section 4 reports the case study, that is, the application of GMERF to Politecnico di Milano data to predict students dropout probability and in Section 5 we draw our conclusions.

## 2 | METHODS

In this section, we recall the basics of generalized linear mixed-effects models (Section 2.1) and we describe the proposed GMERF model, together with the algorithm for the estimation of its parameters (Section 2.2).

### 2.1 | Generalized linear mixed-effects models

Let consider a generalized linear mixed-effects model (GLMM), described in [32]. GLMM is an extension of the generalized linear model (GLM) [28] that includes both fixed and random effects in the linear predictor. GLMMs handle a wide range of response distributions where observations have a hierarchical structure, that is, they are grouped at different levels. Therefore, the GLMs' independence assumption of the observations is no more valid.

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  $\underline{y}_{-i} = (y_{i1}, \dots, y_{in_i})$  be the  $n_i$ -dimensional response vector for observations in the  $i$ th group. Conditionally on random effects denoted by  $\underline{b}_{-i}$ , a GLMM assumes that the elements of  $\underline{y}_{-i}$  are independent, with density function  $f_i$  from the exponential family, of the form

$$f_i(y_{ij}|\underline{b}_{-i}) = \exp \left\{ \frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} + c(y_{ij}, \phi) \right\},$$

where  $a$  and  $c$  are specified functions,  $\eta$  is the natural parameter, and  $\phi$  is the dispersion parameter. In addition, we have

$$E[y_{ij}|\underline{b}_{-i}] = a'(\eta_{ij}) = \mu_{ij},$$

$$\text{Var}[y_{ij}|\underline{b}_{-i}] = \psi a''(\eta_{ij}).$$

A monotonic, differentiable link function  $g$  specifies the function of the mean that the model equates to the systematic component. Usually, the canonical link

function is used, that is,  $g = (a')^{-1}$ . From now on, without loss of generality, the canonical link function is used. In this case, the model takes the following form:

$$\begin{aligned}\underline{\mu}_i &= E[y_i | \underline{b}_i] \quad i = 1, \dots, I \\ g(\underline{\mu}_i) &= \underline{\eta}_i \\ \underline{\eta}_i &= X_i \beta + Z_i \underline{b}_i \\ \underline{b}_i &\sim \mathcal{N}_Q(0, \Psi),\end{aligned}\quad (1)$$

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  $\sum_{i=1}^I n_i = J$ .  $\underline{\eta}_i$  is the  $n_i$ -dimensional linear predictor vector, where  $X_i$  is the  $n_i \times (P+1)$  matrix of fixed effects regressors (including 1 for the intercept) of observations in group  $i$ ,  $\beta$  is the  $(P+1)$ -dimensional vector of their coefficients,  $Z_i$  is the  $n_i \times (Q+1)$  matrix of regressors for the random effects (including 1 for the random intercept),  $\underline{b}_i$  is the  $(Q+1)$ -dimensional vector of their coefficients, and  $\Psi$  is the  $(Q+1) \times (Q+1)$  within-group covariance matrix of the random effects. Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters. This modeling takes into account the dependence structure among data (i.e., observations are not assumed independent but their nested structure is considered). Mixed-effects models consist in iterative methods that alternate the fixed effects estimates to the random effects ones, disentangling the effects given to the grouping level (i.e., given to the fact that observations are naturally nested within groups) from the others. In particular, the unexplained variability of data is composed by the residual variability plus the variability due to the nested structure of the observations (that corresponds to the variance of the random effects). In order to quantify these variances, the variance partition coefficient (VPC) is a measure of the intraclass correlation introduced in [17] and it is equal to the percentage of variation that is found at the highest level of a hierarchical model over the total variance. It is defined as

$$\text{VPC} = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{lat}^2}, \quad (2)$$

where  $\sigma_m^2$  is the estimated variance of random effects, and  $\sigma_{lat}^2$  is the residual variability that can neither be explained by fixed effects, nor through the group features that are represented by the random effects.

GLMMs parameters are estimated through maximum likelihood or restricted maximum likelihood (REML) methods, as described in [30]. Such estimation methods, for models of this type, do not have closed form solutions,

but optimal parameters are found numerically—for example, with Gaussian quadrature or PQL [33]—in order to estimate the integrals to evaluate the likelihood, which is then maximized through an iterative method.

## 2.2 | Generalized mixed-effects random forest

Our proposed GMERF embeds the use of tree-based ensemble methods within the mixed-effects models structure, for different classes of response variables in the exponential family.

We basically relax the linear assumption on the fixed effects part of a GLMM, by substituting it with a tree-based structure, making the model more flexible and adaptable to different and unknown functional forms. The matrix formulation of the GMERF model takes the following form:

$$\begin{aligned}\underline{\mu}_i &= E[y_i | \underline{b}_i] \quad i = 1, \dots, I \\ g(\underline{\mu}_i) &= \underline{\eta}_i \\ \underline{\eta}_i &= f(X_i) + Z_i \underline{b}_i \\ \underline{b}_i &\sim \mathcal{N}_Q(0, \Psi)\end{aligned}\quad (3)$$

with the same notation of Equation (1).

The fixed effects part  $f(X_i)$  is not assumed to be linear any more, but it is assumed to have a complex and unknown structure that we estimate by a tree-based ensemble method, the RF [9]. The basic idea of a RF is to train a large number of trees, each one using a different dataset built from the original one by bootstrap, and including among the covariates used in each tree only a (random) subset of the  $P$  available ones. The prediction of the RF is a suitable aggregation of the predictions of the built trees.

As in a GLMM,  $\underline{b}_i$  and  $\underline{b}_{i'}$  are assumed independent for  $i \neq i'$ . Fixed effects are identified by a nonparametric RF model associated to the entire population, while random ones are identified by group-specific parameters.

To implement GMERF model, we need to decouple the estimation of fixed and random effects parts, alternating them until convergence. To this purpose, we note that, if random effects were known, the GMERF model implies that we could fit a RF to estimate  $f$  using  $\eta_{ij} - \underline{Z}_{ij}^T \underline{b}_i$  as dependent variable. Similarly, if the population-level effects  $f$  were known, then we could estimate the random effects using a traditional mixed-effects linear model with response corresponding to  $\eta_{ij} - f(\underline{X}_{ij})$ . As neither the random effects nor the fixed effects are known, we implement



an iterative method that alternates, until convergence, the estimation of the RF, relative to the fixed effects part, with the estimation of the random effects. The convergence is reached when the difference between the random effects estimates at two consecutive iterations is lower than a fixed tolerance. A second important aspect to be faced is that  $\eta_{ij}$  is not known and it cannot be directly deduced from data. The solution that we adopt, in line with the one proposed in [15], is estimating it by means of a standard GLM model using as covariates the fixed effects covariates. The pseudo-code of the estimation procedure is shown in Algorithm 1.

In the literature, there has been little exploration of the statistical properties of RFs and of mathematical forces driving the algorithm. Most theoretical studies have concentrated on isolated part or stylized versions of the algorithm [8, 26, 27]. In [7], the authors offer an in-depth analysis of a RF model that is very close to the original one [9], proving that the procedure is consistent and that its rate of convergence depends only on the number of the *strong* features and not on how many noise variables are present. Nevertheless, the statistical mechanism of RFs is not yet fully understood and is under active investigation. Both in the simulation study and in the case study, GMERF algorithm has no trouble to converge. With respect to standard trees for nested data, replacing a single tree by a RF probably helps to stabilize the process. However, an in-depth study of the convergence issue in future work would be useful to better understand the behavior of the GMERF algorithm.

The RF is fitted using the R package *randomForest* [25] which implements the original algorithm described in [9]. The GLMM is fitted using the function *glmer* from the R package *lme4* [5]. To predict a new observation  $[x_{ij}; z_{ij}]$  we use the formula

$$\hat{\eta}_{ij} = \hat{f}(x_{ij}) + z_{ij}^T \hat{b}_i, \quad (4)$$

where  $\hat{f}$  is the RF estimated by the algorithm,  $\hat{b}_i$  is the vector of the random effects coefficients related to the  $i$ th group. The prediction of  $\hat{\mu}_{ij}$  is obtained by applying to the corresponding  $\hat{\eta}_{ij}$  the inverse link function  $g^{-1}$ . For example, for a binary response variable, we use the canonical link function  $\text{logit}(x) = \log(x/(1-x))$ .

### 3 | SIMULATION STUDY

In this section, we perform a simulation study, comparing GMERF performance to other similar classification methods on different simulated datasets, with the aim of evaluating GMERF strengths and weaknesses.

#### 3.1 | Simulation design

Without loss of generality, we simulate the response variable from a Bernoulli distribution.<sup>1</sup> The data generating process (DGP) of binary data is based on the following equations:

$$\begin{aligned} \eta_{ij} &= f(\underline{X}_{ij}) + \sum_{q=1}^Q b_{iq} z_{ijq} \\ \mu_{ij} &= \text{logit}^{-1}(\eta_{ij}) \\ y_{ij} &\sim \text{Bernoulli}(\mu_{ij}), \end{aligned} \quad (5)$$

where  $f$  identifies the fixed effect unknown functional form,  $\underline{X}_{ij}$  is the  $P$ -dimensional vector of fixed effects covariates and  $\sum_{q=1}^Q b_{iq} z_{ijq}$  is the linear random effects part of the model. As far as the fixed effects part is concerned, we consider a sizeable (but not too large) number  $P$  of covariates and we design  $f$  to include both a linear part and a tree-like part, as well as interactions among covariates. In this way, we simulate the case of a very diverse structure that will test the flexibility and adaptability of our method. In particular, we set  $P = 7$  and we design  $f$  in the following way:

$$f(x_1, \dots, x_7) = \alpha(x_1^2 - 3x_2 - x_2x_3^2) + \beta \text{tree}(x_4, x_5, x_6), \quad (6)$$

where  $\alpha$  and  $\beta$  are two parameters used to control the variability of  $f$ ;  $\text{tree}(x_4, x_5, x_6)$  is a function with a tree-like structure, described in Figure 1. The last variable  $X_7$  is no significant by construction and it is included in order to test whether the algorithm is misled by it. The seven covariates are randomly generated according to the following distributions:  $X_1, X_2 \sim U(-1, 1)$ ;  $X_3 \sim \text{Weibull}(3)$ ;  $X_4 \sim U(-3, 3)$ ;  $X_5 \sim U(-6, 6)$ ;  $X_6 \sim U(-5, 5)$ ;  $X_7 \sim U(-4, 4)$ .

Regarding the random effects part, we generate  $N = 10$  groups, each one with  $n_i = 40$  observations<sup>2</sup> (for a total of 400 units) by sampling from a normal distribution, according to the assumption of the GLMM. Regarding the random effects specification, we simulate two different cases:

- *Random intercept only*:  $\sum_{q=1}^Q b_{iq} z_{ijq} = b_{i0} \sim \mathcal{N}(0, \gamma^2)$ , that is, there is one scalar random effect, where  $\gamma$  regulates the variability of the random effect;

<sup>1</sup>Here, we make this choice to be in line with the case study in which the response variable is binary. We recall that the model can deal with any response variable in the exponential family, that is handled by the function *glmer* from the R package *lme4* (since the *glmer* function is a step on the GMERF algorithm), by setting the appropriate link function.

<sup>2</sup>Without loss of generality and for the sake of simplicity, we fix the same size for all groups. Nonetheless, the number of observations is allowed to differ across groups.

---

**Algorithm 1** GMERF model estimation procedure
 

---

**Input:**

$y$ - vector with responses  $y_{ij}$   
 $cov$ - data frame with all covariates  
 $gr$ - vector with the grouping variable for each observation  
 $zname$ - vector with names of covariates to be used as random effects  
 $xname$ - vector with names of covariates to be used as fixed effects  
 $fam$ - distribution of  $y$  (must be part of the exponential family)  
 $b_0$ - optional matrix of initial values for each  $\underline{b}_i$   
 $toll$  threshold to decide whether our estimation converged or not  
 $itmax$  maximum number of iterations

$Z \leftarrow (1; cov[zname])$  {to include also the random intercept}

Initialize  $b$  to a matrix of zero (if  $b_0$  is not given) {Each column  $b[i, ]$  of  $b$  will be the  $i$ -th random coefficients  $\underline{b}_i$ }

$all.b[0] = b$

fit a GLM model using  $y$  as response and  $cov$  as matrix of covariates

$eta \leftarrow$  estimated  $\eta_{ij}$  by the GLM model

$it \leftarrow 1$

**while**  $it < itmax$  **and not**  $conv$  **do**

$targ \leftarrow eta - Z \times b$

    fit a random forest model using  $targ$  as target and  $cov$  as predictor matrix

$fx \leftarrow$  fitted values of the forest model

    fit the GLMM  $\eta_{ij} - f(\underline{x}_{ij}) = \underline{z}_{ij}^T \times \underline{b}_i$

$all.b[it] \leftarrow b \leftarrow$  the estimated  $b$  from the model

$M \leftarrow \max(abs(b - all.b[it - 1]))$

$(i, j) \leftarrow \operatorname{argmax}(abs(b - all.b[it - 1]))$

$tr \leftarrow M / all.b[it - 1](i, j)$

**if**  $tr < toll$  **then**

$conv \leftarrow \text{true}$

**else**

$conv \leftarrow \text{false}$

**end if**

$it++$

**end while**

**if not**  $conv$  **then**

    give a warning

**end if**

**Output:**

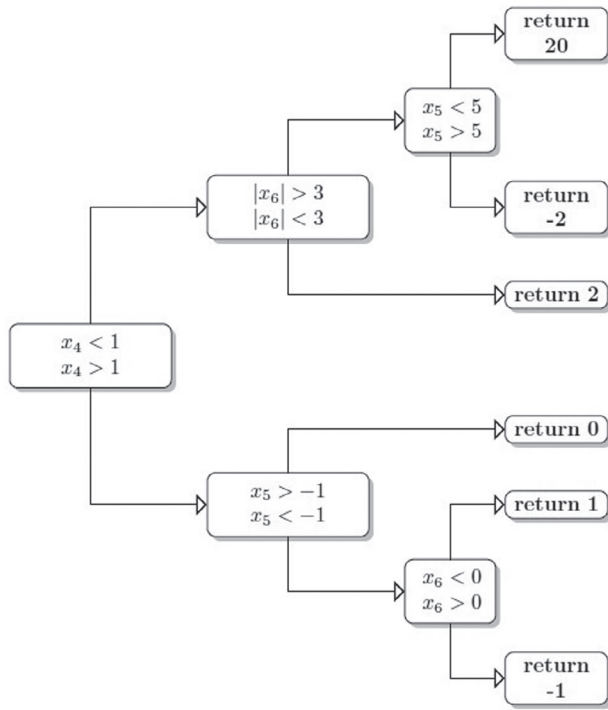
    the final GLMM fitted

    the final forest model fitted

$b$ , the final estimation of the random coefficients

$it$ , the number of iterations

---



**FIGURE 1** Tree-like part ( $tree(x_4, x_5, x_6)$ ) of the fixed effects structure in Equation (6)

- **Random intercept and slope:**  $\sum_{q=1}^Q b_{iq}z_{ijq} = b_{i0} + b_{i1}x_{ij1}$ , where  $x_{1ij}$  is the first fixed effects covariate and the random coefficient is  $\underline{b} \sim \mathcal{N}_2(0, \Sigma)$ , with  $\Sigma = \text{diag}(\gamma^2; \delta^2)$ .  $b_{0i}$  and  $b_{1i}$  are independent for any value of  $i$ ;  $\delta$  is a variance-regulation parameter as well. In this case,  $x$  is a covariate whose effect is not assumed as fixed for all observations but is group-specific, that is, we assume  $x$  to have different effects across observations belonging to different groups.

The presented parameters regulate the variability in the simulated data; we select their values in order to generate probability  $\mu_{ij}$  of each unit not too close to 0 or 1 (except for a small number of observations). We perform a total of eight simulation cases in which we change the value of each coefficient in order to simulate the case of low or high variance for the corresponding component of the model; the chosen coefficients values are summarized in Table 2.

We compare GMERF's performance with the ones of the following methods: GLM; GLMER, which fits the GLMM and is part of the R package *lme4* [5]; RF, which can be fitted using the R package *randomForest* [25]; GMET, described in [15]; support vector machines (SVMs), implemented in the R package *e1071* [37] and the GLMER tree from the *glmertree* R package [14]. We select these methods to—at least partially—cover the panorama of both generalized linear and tree-based mixed-effects models and the most recent classification techniques.

### 3.2 | Simulation results

For each of the eight combinations of fixed and random effects parameters described in Table 2 and each of the seven models, we simulate the dataset 100 times and we analyze the distribution of the results. In order to evaluate the predictive performances, we generate, together with each training dataset, a test dataset, consisting of 50 observations for each group (a total of 500 observations).

To evaluate the quality of the predictions we compute two indexes: predictive mean absolute deviation (PMAD) and predictive misclassification rate (PMCR), which are defined as

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

$$\text{PMCR} = \frac{1}{N_{\text{test}}} \sum_{i=1}^I \sum_{j=1}^{n_i} |y_{ij} - \hat{y}_{ij}|, \quad (7)$$

where  $N_{\text{test}} = 500$  and  $n_i = 50 \forall i = 1, \dots, 10$ ;  $\mu_{ij}$  is the actual probability simulated by the DGP in (5),  $\hat{\mu}_{ij}$  is the probability predicted by the model,  $y_{ij}$  is the actual value of the response and  $\hat{y}_{ij}$  is the response predicted by the model. Simulation results are shown in Table 3. Note that RF and SVM algorithms do not produce probabilities as output, but just the actual responses, so PMAD is not available for them.

The predictive performances of the compared methods are good, being the average PMAD slightly greater than 0.1, overall, while the misclassified samples are roughly 1 every 5.

The best mean performances, both in terms of PMAD and PMCR, are the ones of GLM and GLMER models. GMET model always performs worse than those, while GMERF is mostly comparable to them. On average, RF performs worse, except for the cases in which fixed effects variance is large and random effects one is small. GLMERtree results to be the worst performing algorithm, which constantly has PMAD and PMCR 0.1 higher than the ones of the other algorithms (besides having high performance variance). As for the SVM method, it performs better when the random effects variability increases, reaching an accuracy close to the one of the best performing models; in this sense it is complementary to the RF, which performs better on the other simulation cases. The two cases in which RF have the best performances are the ones with large-variability fixed effects and small-variability random effects. This result confirms that RF is very efficient at identifying fixed effects, but, when the hierarchical structure is relevant, mixed-effects models, that take into account the hierarchy, over-perform it. In this sense, GMERF performances follow the RF ones -



**TABLE 2** Simulation parameters of both fixed effects (Equation (6)) and random effects parts for the simulation data process

Random effects	Fixed effects variability	$\alpha$	$\beta$	Random effects variability	$\gamma^2$	$\delta^2$
Intercept only	Small	0.4	0.25	Small	0.5	0
Intercept only	High	0.7	0.6	Small	0.5	0
Intercept only	Small	0.4	0.25	High	2	0
Intercept only	High	0.7	0.6	High	2	0
Intercept and slope	Small	0.4	0.25	Small	0.3	0.5
Intercept and slope	High	0.7	0.6	Small	0.3	0.5
Intercept and slope	Small	0.4	0.25	High	1.4	1.4
Intercept and slope	High	0.7	0.6	High	1.4	1.4

having the worst values of PMCR in correspondence of the worst PMCR RF values—and they are better than RF ones when the hierarchy is not negligible.

The performances of all algorithms are overall comparable, especially in the average PMAD value, which almost never differs more than 0.02 between two different algorithms.

As for the variances in the estimations, GMERF represents a big improvement with respect to GMET: GMET is often the one having the largest variance (especially for PMAD), whereas GMERF is the one with the smallest variance (especially in PMAD, where this happens six times out of eight). This is a big upside of the algorithm, which proves to provide more stable estimates; this is probably due to the iterative nature of the algorithm, which stabilizes the estimates. GMERF algorithm iterates between estimations of a GLMER model and a RF one; this makes GMERF to get the advantages of both algorithms and justifies the low variability: if one of the two performs bad on a given dataset the other one can compensate; thus, the variability of the misclassification can be kept under control. This justifies the improvement in the algorithm obtained by replacing the tree estimate with a forest.

In conclusion, GMERF algorithm performs comparably to GLMER and GLM, particularly when fixed effects are larger than random effects, but its estimates result to be more stable; this can be seen in the same way as ridge regression versus classical linear regression: ridge is biased, but its estimates have lower variance and, in some cases, it is preferable to its unbiased alternative.

## 4 | CASE STUDY

In this section, we present a real life application of GMERF method: we give our contribution to the SPEET project by using GMERF to model and predict students dropout,

as introduced in Section 1. The aim of this study is to apply GMERF to university students data in order to predict the student dropout probability considering students information—including demographics, previous studies and performances at the beginning of their academic career—and the engineering degree programs they are enrolled in. In particular, we review the case study proposed in [15], where the authors apply GMET to the same dataset: in our application we compare GMERF results with GMET ones.

### 4.1 | The dataset

The data come from Politecnico di Milano database, that consists of 41,098 engineering careers in Bachelor of Science (BSc) that began between A.Y. 2010/2011 and 2015/2016. Politecnico di Milano has  $I = 23$  different engineering degree programs and students are structurally nested within those programs. A descriptive analysis shows that a high percentage of students (27%, more than one out of four) leaves Politecnico di Milano before obtaining the degree. Our goal is to find out which student-level indicators are able to discriminate between two different profiles: *dropout* and *graduate* students. Standing on previous literature [2], there are typically three macro-areas of student-level information that result to be significant in predicting student dropout: student collateral data (i.e., general personal information about the student), student previous studies (i.e., information about the studies of the student before enrolling at the university), student career data (i.e., track of the career of the student at the university, including exams, scores, and mobilities). Taking this prior knowledge into account and after some explorative analysis, we select from Politecnico di Milano database, as variables to be included in the model, the student information that we think could be more informative: personal

**TABLE 3** Prediction performances of the seven methods in each of the eight simulation cases listed in Table 2

Model setting	Feff var	Reff var	Algorithm	Mean PMAD	Var PMAD	Mean PMCR	Var PMCR
Int	Small	Small	GLM	0.112	0.14	0.365	0.694
Int	Small	Small	GLMER	0.11	0.118	0.363	0.769
Int	Small	Small	RF	—	—	0.372	0.703
Int	Small	Small	GMET	0.116	0.141	0.369	0.922
Int	Small	Small	GMERF	0.111	0.102	0.36	0.718
Int	Small	Small	SVM	—	—	0.378	0.732
Int	Small	Small	GLMERTree	0.141	0.197	0.383	0.963
Int	Large	Small	GLM	0.166	0.075	0.31	0.473
Int	Large	Small	GLMER	0.165	0.064	0.308	0.491
Int	Large	Small	RF	—	—	0.296	0.486
Int	Large	Small	GMET	0.172	0.065	0.31	0.417
Int	Large	Small	GMERF	0.172	0.065	0.308	0.511
Int	Large	Small	SVM	—	—	0.32	0.484
Int	Large	Small	GLMERTree	0.185	0.089	0.322	0.438
Int	Small	Large	GLM	0.089	0.168	0.239	2.012
Int	Small	Large	GLMER	0.09	0.183	0.239	2.107
Int	Small	Large	RF	—	—	0.283	2.641
Int	Small	Large	GMET	0.097	0.201	0.243	2.059
Int	Small	Large	GMERF	0.096	0.108	0.241	2.404
Int	Small	Large	SVM	—	—	0.242	2.445
Int	Small	Large	GLMERTree	0.26	1.569	0.394	3.101
Int	Large	Large	GLM	0.133	0.196	0.23	1.1
Int	Large	Large	GLMER	0.135	0.208	0.23	1.09
Int	Large	Large	RF	—	—	0.269	1.271
Int	Large	Large	GMET	0.141	0.212	0.236	1.198
Int	Large	Large	GMERF	0.142	0.168	0.238	1.126
Int	Large	Large	SVM	—	—	0.238	1.063
Int	Large	Large	GLMERTree	0.264	1.279	0.357	1.84
Int + Slope	Small	Small	GLM	0.111	0.107	0.351	0.559
Int + Slope	Small	Small	GLMER	0.111	0.122	0.351	0.555
Int + Slope	Small	Small	RF	—	—	0.369	0.727
Int + Slope	Small	Small	GMET	0.117	0.19	0.358	0.704
Int + Slope	Small	Small	GMERF	0.111	0.092	0.352	0.508
Int + Slope	Small	Small	SVM	—	—	0.366	0.758
Int + Slope	Small	Small	GLMERTree	0.15	0.258	0.381	0.746
Int + Slope	Large	Small	GLM	0.164	0.057	0.307	0.483
Int + Slope	Large	Small	GLMER	0.164	0.051	0.305	0.381
Int + Slope	Large	Small	RF	—	—	0.293	0.412

TABLE 3 Continued

Model setting	Feff var	Reff var	Algorithm	Mean PMAD	Var PMAD	Mean PMCR	Var PMCR
Int + Slope	Large	Small	GMET	0.172	0.049	0.31	0.483
Int + Slope	Large	Small	GMERF	0.172	0.048	0.306	0.523
Int + Slope	Large	Small	SVM	—	—	0.316	0.499
Int + Slope	Large	Small	GLMERTree	0.187	0.124	0.321	0.714
Int + Slope	Small	Large	GLM	0.092	0.168	0.242	1.702
Int + Slope	Small	Large	GLMER	0.094	0.19	0.241	1.697
Int + Slope	Small	Large	RF	—	—	0.288	1.908
Int + Slope	Small	Large	GMET	0.101	0.22	0.247	1.859
Int + Slope	Small	Large	GMERF	0.099	0.138	0.243	1.735
Int + Slope	Small	Large	SVM	—	—	0.245	2.026
Int + Slope	Small	Large	GLMERTree	0.256	1.53	0.396	3.053
Int + Slope	Small	Large	GLM	0.134	0.248	0.24	1.408
Int + Slope	Small	Large	GLMER	0.137	0.249	0.241	1.409
Int + Slope	Small	Large	RF	—	—	0.275	0.993
Int + Slope	Small	Large	GMET	0.143	0.272	0.242	1.346
Int + Slope	Small	Large	GMERF	0.145	0.225	0.242	1.33
Int + Slope	Small	Large	SVM	—	—	0.246	1.36
Int + Slope	Small	Large	GLMERTree	0.265	0.928	0.359	2.034

information, previous studies and the career track at the first semester of the first year, in the perspective of predicting the student dropout probability as soon as possible ([6, 16]). Table 4 reports the selected variables, with their description. As students are naturally nested within degree programs, we include in the model a random intercept given to the degree program in which students are enrolled in, in order to take into account this source of dependence among students and to investigate possible differences in the dropout phenomenon across degree programs.

We exclude from the dataset four degree programs having few students enrolled (less than 200), so the final number of degree programs considered is  $I = 19$ . The statistical units are represented by the concluded (either graduated or dropout) careers of students enrolled in the 19 selected degree programs.<sup>3</sup> The final dataset regards 24,736 students (statistical units) nested within 19 degree programs.<sup>4</sup>

<sup>3</sup>The 19 engineering degree programs are: Aerospace, Automation, Biomedical, Building, Chemical, Civil, Civil and Environmental, Electrical, Electronic, Energy, Computing Systems, Environmental and Land Planning, Industrial Production, Management, Materials and Nanotechnology, Mathematical, Mechanical, Physics, Telecommunications.

<sup>4</sup>This dataset coincides with the one used in [15], with the only difference that variable *Previous studies* here has four levels, whereas in

We randomly split the dataset into training and test sets, with a ratio of 70% for model fitting and 30% for evaluation (which we will refer to as test set). We then split again the model fitting set into a training set and a validation set using a proportion of, respectively, 80% and 20%; validation set will be used to select the threshold value  $p_{opt}$  to classify students.

## 4.2 | Model results

The mixed-effects model we implement considers *Status* as binary response variable; all variables from *Sex* to *Credits* in Table 4 as fixed effects covariates and it includes a random intercept  $b_0$  for the degree program. We apply GMERF model using  $toll = 0.02$  and  $itmax = 30$ . It converges after eight iterations (computational time = 2.33 min), proving to reach the stability quite quickly. Estimates of random intercepts together with their confidence intervals are shown in Figure 2. We notice that 10 intercepts are not significantly different from 0 (with 95% confidence), being in line with the average. Five programs increase the log-odds of dropout,

[15] it has only three levels (“Classica” and “Altro” are considered as a unique level); this has a minor impact on the final results of the analysis, so a comparison between the two methods is still possible.

**TABLE 4** Politecnico di Milano student-level variables

Variable name	Type of variable	Domain	Description
<i>Status</i>	Factor	{0, 1}	Binary response variable about the status of the concluded career: 1 = Dropout, 0 = Graduate
<i>Sex</i>	Factor	{"M", "F"}	Gender of the student
<i>Nationality</i>	Factor	{"I", "F"}	Nationality of the student: "I" = Italian, "F" = Foreign
<i>Previous studies</i>	Factor	Four levels	Type of studies before university (secondary school): "Scientifica," "Classica," "Tecnica," "Other"
<i>Avg Score</i>	Numeric	$\{0\} \cup [18; 30]$	Weighted average score obtained in exams of the first semester of the first year
<i>Attempts</i>	Numeric	[0; 5]	Average number of attempts per exam of the first semester of the first year
<i>Credits</i>	Integer	{1, ..., 50}	Total number of <i>Crediti Formativi Universitari</i> (CFU) obtained by the student after the first semester of the first year.
<i>Degree program</i>	Factor	23 levels	Degree program the student is enrolled in (grouping variable)

Notes: Variable *Avg Score* has a peculiarity, in the sense that it takes values from 18 to 30 (the minimum and maximum possible score) plus a point mass at 0, representing students who did not pass any exam.

while four programs decrease them. As the variance of the standard logistic distribution is  $\pi^2/3 \simeq 3.29$ , the VPC can be estimated as:

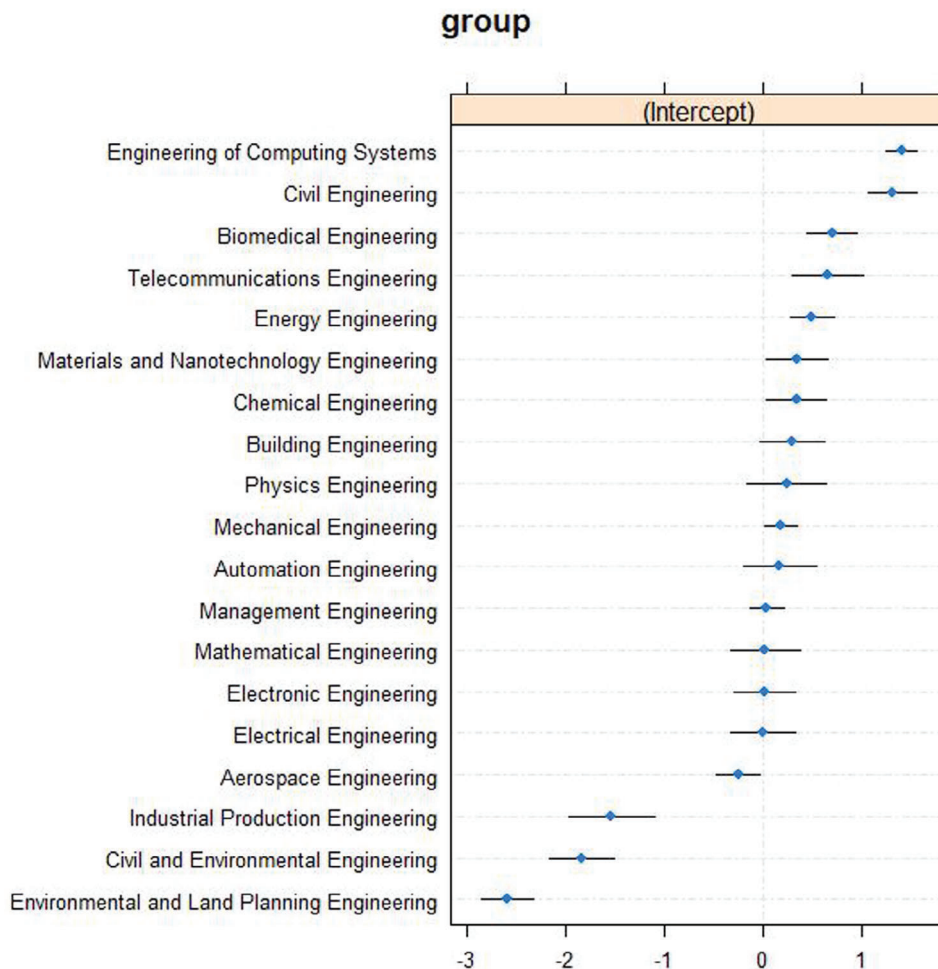
$$\text{VPC} = \frac{\sigma_m^2}{\sigma_m^2 + \pi^2/3} = 0.2261.$$

Roughly 23% of unexplained variation in the response is attributable to the nested structure of students. This value of VPC, together with the fact that some random intercepts are significantly different from zero (Figure 2) suggests that there is a relevant heterogeneity in the dropout phenomenon across degree programs and highlight the importance of taking into account the hierarchical structure of these data. Regarding the fixed effects part, RF model measures the importance of each covariate (measured as the increase of the residual sum of squares when the values of the corresponding variable are randomly permuted in the training dataset) in explaining the response and the partial effect of each covariate (that can be displayed by means of partial dependence plots). Figure 3 reports the variables importance plot. GMERF model identifies *Avg 1.1* and *CFU 1.1* as the most important variables. In particular the three covariates associated with performance of the student during his/her first semester of career are more important than student collateral information; this suggests that the choice of leaving the studies is mainly associated to the university performance of the student, more than his/her background when enrolling. This variables importance measure method relies on the choice of a performance measure and is well known and frequently adopted. Nonetheless, a very recent branch of the literature proposes an alternative method that selects the variables

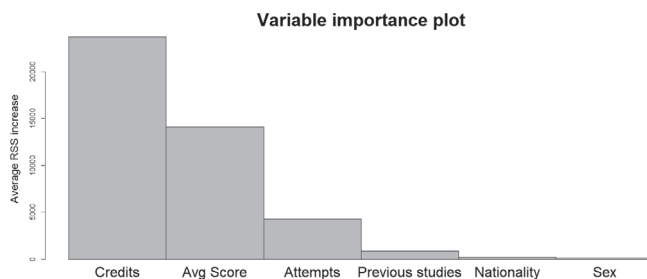
depending on the structure of the trees [22, 31, 13]. In particular, the authors in [22] propose an algorithm based on the minimal depth (MD) statistic, that is, based on the idea that variables that tend to split close to the root node should have more importance in prediction. By removing the dependence on performance measures, the arrangement of the trees gains strength, as in the case of splitting rules. Equivalently, the authors in [31] propose a new alternative importance measure, called intervention in prediction measure (IPM), that follows the same approach of MD but, unlike MD, is a case-based method. In order to check the robustness of variables importance in Figure 3, we measure the IPM variables importance, that is expressed as a percentage. Table 5 and Figure 4 report numerical and graphical results, respectively.<sup>5</sup> The rankings of the variables, standing on their estimated importance, obtained by the two methods coincide.

Variables importance plot identifies the important covariates but it does not give insights about the type of association between the covariates and the response (e.g., whether they are directly or inversely related to the outcome or the relevant covariate values range). Partial plots highlight the association of each covariate with the response, net to the other covariates included in the model. Figures 5 and 6 show the partial plots for all continuous and categorical fixed effects covariates, respectively. In particular, for the variable *Avg Score*, we show two different

<sup>5</sup>As IPM method does not support categorical covariates (with more than two categories), we recoded the *PreviousStudies* covariate as two separate dummy variables: *Clas* (Classica vs. Scientifica) and *Tecn* (Tecnica vs. Scientifica).



**FIGURE 2** Random intercepts relative to the 19 degree programs estimated by the GMERF model with their confidence intervals



**FIGURE 3** Plot of GMERF's fixed effects variables importance to predict student dropout; the height of the bar is the increase of the residual sum of squares (RSS) when the values of the corresponding variable are randomly permuted

plots, Figure 5A,B: the former shows the plot with respect to the entire range values of *Avg Score*, while the latter focuses just on the values from 18 to 30; the jump after 0 in the first one is motivated by the fact that there are no values of this variable in the interval (0; 18).

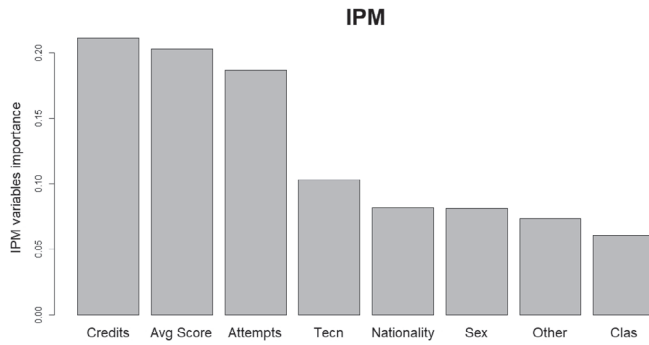
By looking at Figure 5B,D, we notice an inverse proportional association between the dropout probability and the values of variables *Avg Score* and *Attempts*, that suggests that students trying less exams and not passing

**TABLE 5** IPM of the fixed-effects covariates used in the GMERF method to predict student dropout

Variables	IPM
Credits	0.218
Avg Score	0.206
Attempts	0.187
Tecn (vs. scientific)	0.095
Nationality	0.083
Sex	0.081
Other (vs. scientific)	0.073
Class (vs. scientific)	0.057

them at the first semester tend to drop their studies. (In particular, Figure 5B shows that the dropout probability decreases linearly with variable *Avg Score*.) This pattern repeats in Figure 5C, even if not in the same straightforward way. From this figure we also note that students who obtain 30 CFU after the first semester (i.e., the student passes all exams of that semester) has almost null dropout





**FIGURE 4** Intervention in prediction measure (IPM) of the fixed-effects covariates used in the GMERF method to predict student dropout

probability; this strongly suggests that a student likely to dropout can be identified after a semester of studies.

Regarding the previous studies, Figure 6 shows that there is not a significant difference in the dropout probability of students who attended scientific, classic, or other schools (after adjusting for the other characteristics), whereas students who attended technical schools are more likely to dropout.

GMERF model estimates the probability that a student drops his/her studies. To evaluate the quality of the predictions we use four indexes: Accuracy  $A$ , that is the percentage of correctly classified units; Sensibility  $SN$  that is, out of all the positive units, the proportion of those found by the algorithm; Specificity  $SP$  that is, out of all the positive-predicted units, the percentage of those who actually are;  $F1$ -measure, which combines Sensitivity and Specificity as

$$F1 = \frac{2 \cdot SN \cdot SP}{SN + SP}. \quad (8)$$

We use the validation set to choose the optimal threshold value  $p_{opt}$  for prediction, by looking at the prediction accuracy and at the ROC curve that we build with this set [1]. In Figure 7, the complete ROC curve Sensitivity-Specificity is shown. The optimal value turns out to be  $p_{opt} = 0.4$ , both in terms of Accuracy ( $A = 0.9082$ ) and  $F1$ -measure ( $F1 = 0.8305$ ); the other indexes values are  $SN = 0.8102$  and  $SP = 0.8495$ . The relative misclassification table is.

	$y = 0$	$y = 1$
$\hat{y} = 0$	2366	180
$\hat{y} = 1$	138	779

These results show a high predictive power of the model in the validation set. We now test the method on the test set.

The misclassification table of the predictions on the test set is.

	$y = 0$	$y = 1$
$\hat{y} = 0$	5138	406
$\hat{y} = 1$	273	1603

and the value of the indexes are:

$$A = 0.9085$$

$$SP = 0.8544$$

$$SN = 0.7979$$

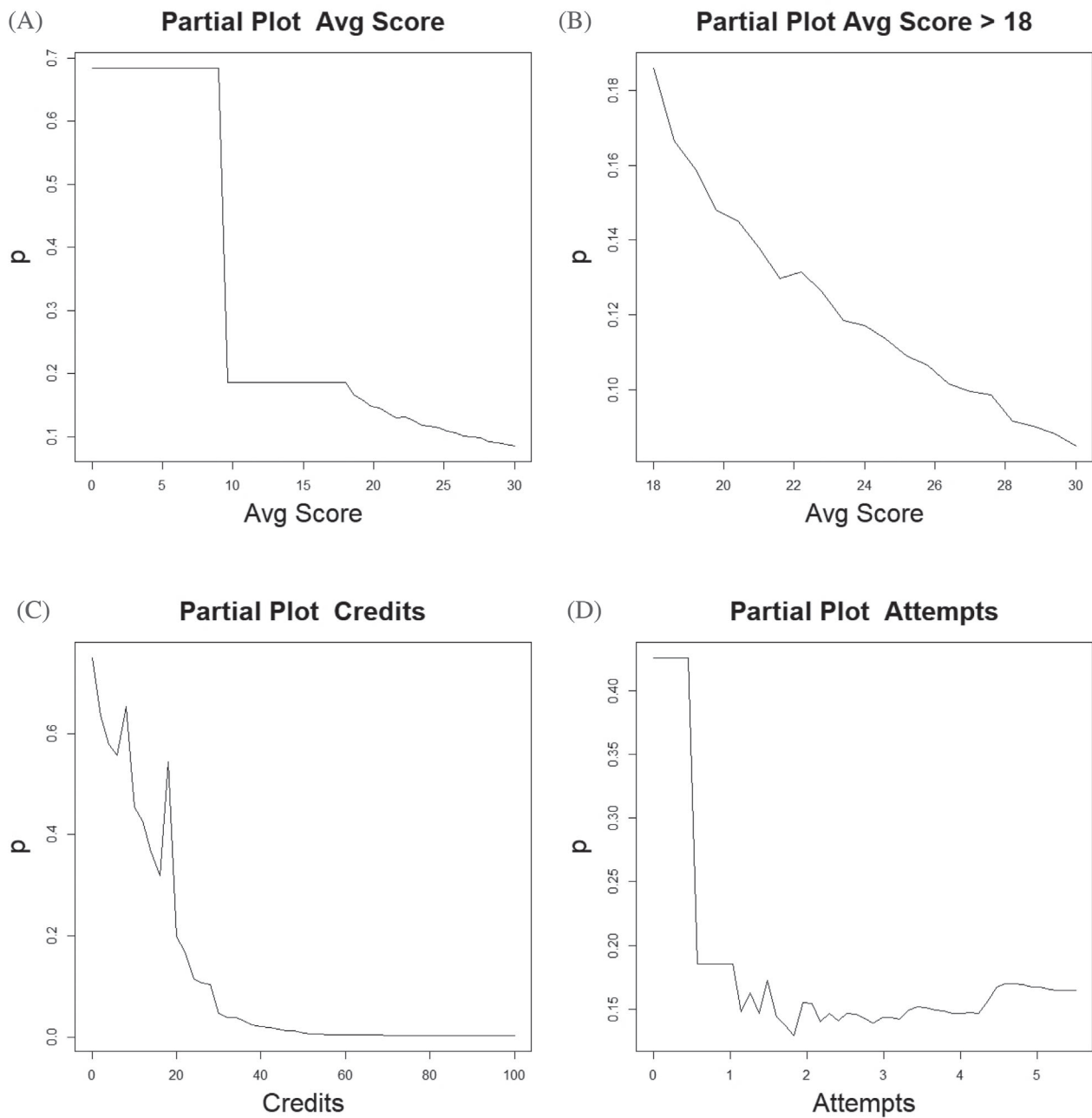
$$F1 = 0.8252.$$

Overall, GMERF model gives the right prediction 91% of times, 80% of students who drop their studies are correctly identified and 85% of students predicted as *dropout* actually are; these high values, especially regarding the indices  $SN$  and  $SP$ , reveal that the method is very accurate, but, at the same time, it is less sensitive in predicting students to drop their studies.

In the perspective of investigating the advantages of RF with respect to simple trees, we now compare our results with the ones found in [15] by using the GMET model on the same dataset. Both models identify variables *Credits* and *Avg Score* as the two most important variables to predict the dropout; on the other end, variable *Sex* is not considered significant by either of them. As far as random effects are concerned, both models identify *Environmental and Land planning* engineering as the one associated with the lowest dropout rate and they also both associate *Computer* and *Civil* engineering with high dropout probabilities.

The major differences between the fixed effects part estimated by GMET and GMERF are the following:

- Variable *Attempts* is considered important by GMERF, but it does not appear in GMET as splitting node; this may happen because the effect of this variable is masked, in GMET, by the first split based on variable *CFU 1.1*; the two variables, at least for very small values, are naturally correlated (people attempting no exams do not pass exams and therefore do not get any CFU); however, the RF used in GMERF uses different variables in different trees and is then able to distinguish the effects of both variables, which is one of the main advantages of RF over classification trees;
- *Nationality* is considered very important by GMET, being the second split, while in GMERF it has almost null importance.



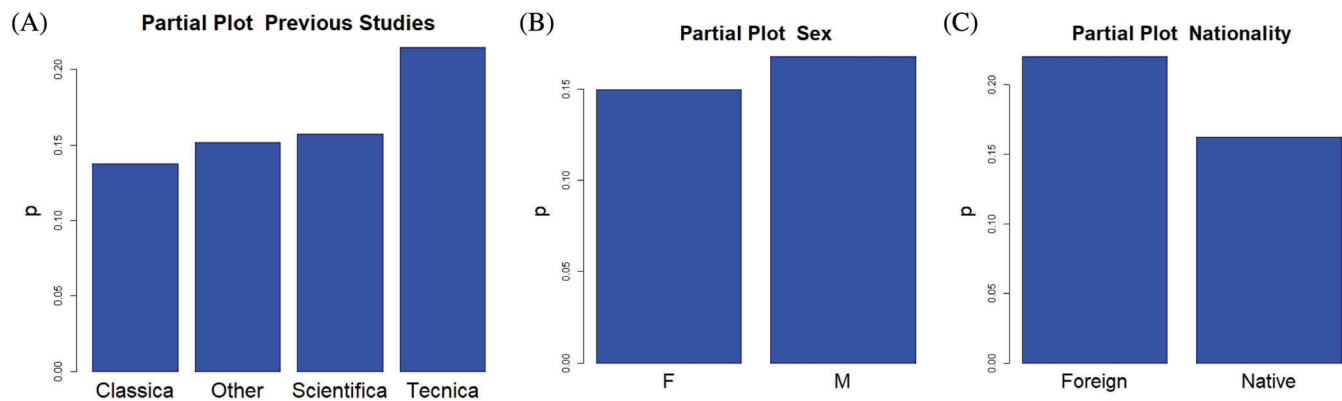
**FIGURE 5** Partial plots of student dropout probability with respect to continuous variables: variable *Avg Score* on the entire range in panel (A), variable *Avg Score* on the range (18; 30) in panel (B); variable *Credits* in panel (C) and variable *Attempts* in panel (D). The y-axis reports the increment/decrement in dropout probability, given to the covariate on the x-axis

Regarding the estimation of random effects, the two compared methods identify a similar trend in the association between the 19 degree programs and students dropout probability, except for:

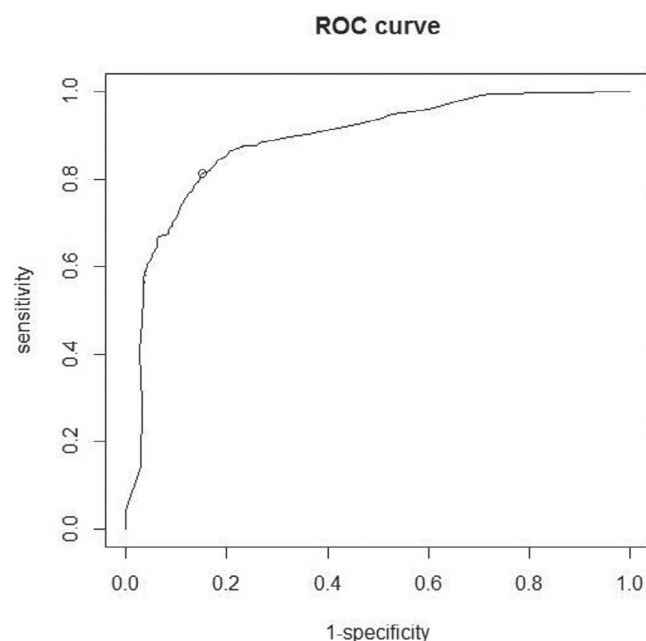
- *Management Engineering*, which our model considers in line with the average, while GMET associates it to a lower dropout probability, with respect to the average;
- *Biomedical Engineering* and *Telecommunications Engineering*, which in our model are associated to a higher

dropout probability, while in GMET they are associated to a null random intercept.

Comparing the predictive power of the two models on the test set, we see that GMERF brings a slight improvement to the accuracy, from GMET's 0.878 to 0.908; therefore 3% more of the students are correctly classified, which confirms our initial expectation. Overall we can say that the two models highlighted similar dynamics, which is an evidence on the robustness of the two of them;



**FIGURE 6** Partial plots of the student dropout probability with respect to categorical variables: *PrevStudies*, *Sex*, and *Nationality*. The y-axis reports the increment/decrement in dropout probability, given to the covariate on the x-axis



**FIGURE 7** ROC curve obtained from the validation set; the point highlighted is the one corresponding to the optimal value of  $p_{opt}$  found with the validation process

the major difference is the higher precision with which GMERF model classified students and showed the effects of each covariate on the dropout probability.

## 5 | CONCLUSION

In this work, we present a method called GMERF, which consists in a novel method that extends the use of RF to the analysis of hierarchical data, for a non-Gaussian response variable. GMERF modeling substitutes the linear combination of the fixed-effects covariates of a GLMM with a RF. This new method contributes to the statistical literature

about mixed-effects models and tree-based method, taking advantage of the flexibility and the predictive power of a RF, but maintaining the structure of mixed-effects models. Although our simulation and case studies focus on the binary response case, this approach can handle any type of response variable in the exponential family. Using suitable link functions, GMERF is able to model different outcomes such as counts data, as well as the particular case of a Gaussian response. GMERF can be considered a step forward in the class of models which combine tree-based methods with linear mixed models. The simulation study shows that GMERF has prediction performances comparable to models like GLM and GLMM, with the advantage that its estimates are less variable than the ones of these models; moreover, it has the added benefit of not assuming any functional form on the fixed effect part and it can deal with heterogeneous covariates (categorical and continuous) at the same time, which is a very big advantage in terms of flexibility. In particular, the RF part is able to model the association between each covariate and the response, net to the effect of all the other covariates, identifying possible nonlinear trends and range-dependent patterns. The advantages of using GMERF, instead of GLMM, depend on the complexity of the fixed effects structure (that, most of the time, is not known a priori): if fixed effects covariates are linearly associated to the response and are not correlated, GLMM is expected to perform better; on the opposite, if the covariates set is substantially large and the covariates potentially interact among each other creating complex patterns in their association to the response, then, we expect GMERF to outperform parametric methods with predefined functional forms. In other words, as the random effects estimates follow the same procedure as in GLMM, the potential of GMERF emerges depending on the fixed effects covariates that describe the different real cases.

In the case study, we give a contribution to the SPEET project, by providing an accurate method to classify students as *dropout* or *graduate* that resulted to be successful in the 90% of cases. These results might be useful in the perspective of defining new tutoring systems to help students at risk. Our study results in an improvement in the prediction accuracy over the GMET model, which was applied on the same dataset; this is one of the goals we expect to achieve when using GMERF, since the two models have the same formulation, but GMERF uses a RF to estimate the fixed effects, which is an algorithms that improves the regression tree used in GMET (besides the fact that GMET is not iterative). GMERF proves to be a powerful and an easily interpretable method that can be applied to various complex real data problem.


## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article.

## ORCID

Chiara Masci  <https://orcid.org/0000-0002-9208-3194>

Francesca Ieva  <https://orcid.org/0000-0003-0165-1983>

Anna M. Paganoni  <https://orcid.org/0000-0002-8253-3630>

## REFERENCES

1. A. Agresti and M. Kateri, *Categorical data analysis*, Springer, Heidelberg, 2011.
2. W. Arulampalam, R. Naylor, and J. Smith, *Factors affecting the probability of first year medical student dropout in the UK: A logistic analysis for the intake cohorts of 1980–92*, Med. Educ. 38 (2004), 492–503.
3. S. Athey, J. Tibshirani, and S. Wager, *Generalized random forests*, Ann. Stat. 47 (2019), 1148–1178.
4. M. Barbu, R. Vilanova, J. Lopez Vicario, M. J. Pereira, P. Alves, M. Podpora, M. Ángel Prada, A. Morán, A. Torreburno, S. Marin, and R. Tocu, Data mining tool for academic data exploitation: Literature review and first architecture proposal. Project SPEET – Student Profile for Enhancing Engineering Tutoring, 2017.
5. D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, F. Scheipl, and G. Grothendieck, Package ‘lme4’, Linear mixed-effects models using S4 classes. R package version, 1-1, 2011.
6. F. Belloc, A. Maruotti, and L. Petrella, *University drop-out: An Italian experience*, High. Educ. 60 (2010), 127–138.
7. G. Biau, *Analysis of a random forests model*, J. Mach. Learn. Res. 13 (2012), 1063–1095.
8. G. Biau and L. Devroye, *On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification*, J. Multivar. Anal. 101 (2010), 2499–2518.
9. L. Breiman, *Random forests*, Mach. Learn. 45 (2001), 5–32.
10. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*, CRC Press, Cambridge, MA, 1984.
11. J.-E. Chen and C.-W. Hsiang, *Causal random forests model using instrumental variable quantile regression*, Econometrics 7 (2019), 49.
12. B. Chiandotto and C. Giusti, *L'abbandono degli studi universitari*, Modelli statistici per l'analisi della transizione università-lavoro 7 (2005), 1–22.
13. I. Epifanio, *Intervention in prediction measure: A new approach to assessing variable importance for random forests*, BMC Bioinform. 18 (2017), 230.
14. M. Fokkema, N. Smits, A. Zeileis, T. Hothorn, and H. Kelderman, *Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees*, Behav. Res. Methods 50 (2018), 2016–2034.
15. L. Fontana, C. Masci, F. Ieva, and A. M. Paganoni, *Performing learning analytics via generalized mixed-effects trees*, Mox-report, 2018.
16. P. Goldschmidt and J. Wang, *When can schools affect dropout behavior? A longitudinal multilevel analysis*, Am. Educ. Res. J. 36 (1999), 715–738.
17. H. Goldstein, W. Browne, and J. Rasbash, *Partitioning variation in multilevel models*, Understand. Statist. 1 (2002), 223–231.
18. A. Hajjem, F. Bellavance, and D. Larocque, *Mixed effects regression trees for clustered data*, Statist. Probab. Lett. 81 (2011), 451–459.
19. A. Hajjem, F. Bellavance, and D. Larocque, *Mixed-effects random forest for clustered data*, J. Stat. Comput. Simul. 84 (2014), 1313–1328.
20. A. Hajjem, D. Larocque, and F. Bellavance, *Generalized mixed effects regression trees*, Statist. Probab. Lett. 126 (2017), 114–118.
21. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, Springer Science & Business Media, New York, NY, 2009.
22. H. Ishwaran, U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer, *High-dimensional variable selection for survival data*, J. Am. Stat. Assoc. 105 (2010), 205–217.
23. G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, Vol 112, Springer, New York, NY, 2013.
24. A. H. Li and J. Bradic, *Censored quantile regression forest*, International Conference on Artificial Intelligence and Statistics, 2020, pp. 2109–2119.
25. A. Liaw and M. Wiener, *Classification and regression by random-forest*, R News 2 (2002), 18–22. <https://CRAN.R-project.org/doc/Rnews/>.
26. Y. Lin and Y. Jeon, *Random forests and adaptive nearest neighbors*, J. Am. Stat. Assoc. 101 (2006), 578–590.
27. N. Meinshausen, *Quantile regression forests*, J. Mach. Learn. Res. 7 (2006), 983–999.
28. J. A. Nelder and R. W. Wedderburn, *Generalized linear models*, J. R. Stat. Soc. Ser. A (Gen.) 135 (1972), 370–384.
29. M. Oprescu, V. Syrgkanis, and Z. S. Wu, *Orthogonal random forest for causal inference*, International Conference on Machine Learning, 2019, pp. 4932–4941.
30. H. D. Patterson and R. Thompson, *Recovery of inter-block information when block sizes are unequal*, Biometrika 58 (1971), 545–554.
31. A. Pierola, I. Epifanio, and S. Alemany, *An ensemble of ordered logistic regression and random forest for child garment size matching*, Comput. Ind. Eng. 101 (2016), 455–465.

32. J. Pinheiro and D. Bates, *Mixed-effects models in S and S-PLUS*, Springer Science & Business Media, New York, NY, 2006.
33. G. Rodríguez, *Multilevel generalized linear models*, in *Handbook of Multilevel Analysis*, Springer, New York, NY, 2008, 335–376.
34. C. Romero and S. Ventura, *Educational data mining: A review of the state of the art*, IEEE Trans. Syst. Man Cybernet. C (Appl. Rev.) 40 (2010), 601–618.
35. R. J. Sela and J. S. Simonoff, *RE-EM trees: A data mining approach for longitudinal and clustered data*, Mach. Learn. 86 (2012), 169–207.
36. J. L. Speiser, B. J. Wolf, D. Chung, C. J. Karvellas, D. G. Koch, and V. L. Durkalski, *BiMM tree: A decision tree method for modeling clustered and longitudinal binary outcomes*, Commun. Statist. Simul. Comput. 49 (2018), 1–20.
37. S. Suthaharan, *Machine learning models and algorithms for big data classification*, Integr. Ser. Inf. Syst 36 (2016), 1–12.

**How to cite this article:** Pellagatti M, Masci C, Ieva F, Paganoni AM. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat Anal Data Min: The ASA Data Sci Journal*. 2021;14:241–257. <https://doi.org/10.1002/sam.11505>