A COMPARATIVE ANALYSIS OF DATA MINING METHODS AND HIERARCHICAL LINEAR MODELING USING PISA 2018 DATA

Wenting Weng[1] and Wen Luo[2]

[1]Krieger School of Arts and Sciences, Johns Hopkins University, Baltimore, USA
[2]Department of Educational Psychology, Texas A&M University, College Station, USA

ABSTRACT

Educational research often encounters clustered data sets, where observations are organized into multilevel units, consisting of lower-level units (individuals) nested within higher-level units (clusters). However, many studies in education utilize tree-based methods like Random Forest without considering the hierarchical structure of the data sets. Neglecting the clustered data structure can result in biased or inaccurate results. To address this issue, this study aimed to conduct a comprehensive survey of three tree-based data mining algorithms and hierarchical linear modeling (HLM). The study utilized the Programme for International Student Assessment (PISA) 2018 data to compare different methods, including non-mixed-effects tree models (e.g., Random Forest) and mixed-effects tree models (e.g., random effects expectation minimization recursive partitioning method, mixed-effects Random Forest), as well as the HLM approach. Based on the findings of this study, mixed-effects Random Forest demonstrated the highest prediction accuracy, while the random effects expectation minimization recursive partitioning method had the lowest prediction accuracy. However, it is important to note that tree-based methods limit deep interpretation of the results. Therefore, further analysis is needed to gain a more comprehensive understanding. In comparison, the HLM approach retains its value in terms of interpretability. Overall, this study offers valuable insights for selecting and utilizing suitable methods when analyzing clustered educational datasets.

KEYWORDS

Data Mining, Clustered Data, Mixed-effects, Random Forest, HLM, Hierarchical Linear Modeling, PISA

1. INTRODUCTION

Clustered or hierarchical data exhibits a multilevel structure where observations are sampled from lower-level units (individuals) nested within higher-level units (clusters). This type of data includes attributes at both the individual and cluster levels, enabling the exploration of variations among individuals within and between clusters. Observations within the same cluster tend to share more similarities than those from different clusters. Considering both similarities and differences across clusters is crucial and can lead to more accurate results in research. Clustered data sets are commonly encountered in educational research, such as the Programme for International Student Assessment (PISA) data, which measures the academic achievements of fifteen-year-old students in reading, mathematics, and science. Scholars have studied PISA data using a clustered structure (e.g., [1], [2]).

In 1984, Breiman et al. [3] introduced tree-based methods called classification and regression trees (CART). CART is a non-parametric approach that can handle large data sets with large number of attributes without requiring preselection. CART is particularly robust in handling

outliers, unlike some traditional statistical methods such as linear regression. However, in certain circumstances (e.g., when observations are modified), CART may produce unstable results, leading to high variability and poor predictive performance [4]. To address the instability issue, Breiman [5] proposed a tree-based ensemble method called Random Forest (RF). RF combines a large number of regression trees with the goal of improving predictions. RF has been successfully applied in educational research to predict students' learning performance (e.g., [6]). However, RF only considers the fixed effects of attributes, even when the data has a clustered structure. To overcome this limitation, a new method called the random effects expectation minimization recursive partitioning method (RE-EM tree) was proposed based on CART by Sela and Simonoff [7]. This method takes into account the random effects within a clustered data structure. Subsequently, another approach called mixed-effects Random Forest (MERF) was introduced, which incorporates random effects into RF [8]. This allows for the consideration of both fixed and random effects of attributes, providing a more comprehensive analysis of clustered data.

This paper aims to conduct a comprehensive survey of various tree-based data mining algorithms and hierarchical linear modeling (HLM), which is one of the most widely used approaches for analyzing clustered educational data sets. The comparative study focuses on comparing non-mixed-effects tree models (i.e., RF) with mixed-effects tree models (i.e., RE-EM tree, MERF), as well as the HLM approach. By evaluating the advantages and disadvantages of each method, this comparison will provide valuable insights for selecting and adopting appropriate methods in the analysis of clustered educational data sets.

In the subsequent sections of the paper, we provide a concise overview of the non-mixed-effects tree-based method (RF), the mixed-effects tree-based methods (RE-EM tree, MERF), and the HLM approach. We then present a comparative study to determine the optimal method by utilizing the PISA 2018 clustered data set. Finally, we report the results obtained and engage in a thorough discussion of the findings.

2. THEORETICAL FRAMEWORK

Educational Data Mining (EDM) is a rapidly growing field that focuses on analyzing data within an educational context using various Data Mining (DM) techniques and tools [25]. Tree-based methods have been commonly employed in educational research. These methods have been utilized in various studies to analyze educational data and gain insights into different aspects of the educational context. For example, Decision Tree has been applied to predict student outcomes such as academic success, dropout risks, and online persistence in web-supported courses (e.g., [26], [27]). Additionally, Random Forest has been utilized in predicting learning performance and detecting instances of online cheating behavior among students [28]. These tree-based methods offer valuable tools for extracting knowledge from educational datasets and facilitating data-driven decision-making in the field of education.

Hierarchical linear modeling (HLM) is widely recognized as the predominant statistical method utilized in educational research, particularly in the analysis of multilevel research data. It has found extensive application in various educational studies, including investigations into the effects of technology usage on student learning achievement [29]. HLM offers a powerful framework for examining the relationships between variables at different levels of analysis, allowing researchers to account for the hierarchical structure of educational data and assess the impact of various factors on student outcomes. Its versatility and capability to handle nested data make it a popular choice for researchers seeking to delve into the complexities of educational phenomena.

2.1. Tree-based Method: Random Forest

Random Forest (RF), introduced by Breiman [9], has gained widespread use in prediction and classification tasks (e.g., [10]), even in scenarios with high-dimensional data [11]. RF is a collection of regression trees that combines the bagging procedure with randomization in variable splitting. Bagging, as proposed by Breiman [5], involves generating random bootstrap samples from the original data. The bootstrap samples are generated by repeatedly drawing from the original data set, with each sample having the same size as the original data. Each tree in the RF is constructed by randomly selecting features from the bootstrap samples. The predictions of RF are determined by averaging the outputs of the individual trees.

The main challenge of RF lies in its interpretability due to the composition of multiple regression trees. However, RF can still provide insights into the relevance of input attributes. When training an RF model, the out-of-bag (OOB) observations are not included in the bootstrap samples. These OOB observations are utilized to evaluate the model's accuracy by calculating the OOB error. This error measure is also helpful in selecting optimal values for tuning parameters, such as the number of randomly selected attributes considered for each split [5].

2.2. Mixed-Effects Methods

2.2.1. Hierarchical Linear Modeling

Hierarchical linear modeling (HLM), or multilevel modeling, is a widely utilized method for analyzing clustered data, which involves nested structures where individuals (lower-level units) are grouped within clusters (higher-level units). This approach is commonly applied in educational research, where individuals are sampled from classes and schools (e.g., [12]). In a two-level model, one level explores the relationships among the lower-level units, while the other level examines how these relationships vary across the higher-level units [13]. For instance, consider a random intercept model, which can be expressed as follows:

$$y_{ij} = \beta_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \quad (1)$$

where:
y_{ij} = response variable value for the individual i nested within the j^{th} cluster unit;
β_{0j} = intercept for the j^{th} cluster unit;
β_1 = regression slope associated with the attribute X_{ij} for the j^{th} cluster unit;
X_{ij} = attribute value of X for the individual i in the j^{th} cluster unit;
ε_{ij} = random error for the individual i in the j^{th} cluster unit.

In the model formula (1), β_{0j} can be written as:

$$\beta_{0j} = \gamma_{00} + U_{0j} \quad (2)$$

where:
γ_{00} = mean intercept across all clustered units, which is a fixed effect;
U_{0j} = a random effect of the j^{th} cluster unit on the intercept.

A combined model can be created using Equation (1) and Equation (2):

$$y_{ij} = \gamma_{00} + U_{0j} + \beta_1 X_{ij} + \varepsilon_{ij} \quad (3)$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$
$$U_{0j} \sim N(0, \sigma_U^2)$$

In this random intercept only model, the parameters are estimated via the variance components σ_ε^2 and σ_{II}^2. σ_ε^2 represents the unexplained variation at the lower level when controlling the attribute X_{it}, while σ_{II}^2 is the unexplained variation at the higher level.

2.2.2. RE-EM Tree

Sela and Simonoff [7] introduced the random effects expectation-maximization recursive partitioning method (RE-EM tree), which is specifically designed to handle clustered and longitudinal data. This method utilizes CART [3] as the underlying regression tree algorithm. In Sela and Simonoff [7], we have sampling individuals or objects i = 1, ..., I at times t = 1, ..., T_i. An observation of an individual for a single time is referred as (i, t). An individual can have multiple observations across different times. For each observation, we have a vector of j attributes, $x_{it} = (x_{it1}, \ldots, x_{itj})'$. The attributes may be constant among individuals over time or differ across time and individuals. To detect differences for individuals over time, we have a known design matrix Q_{it} and a vector of unknown individual-specific random effects intercept w_i being uncorrelated with the attributes. A general effects model can be written as:

$$y_{it} = Q_{it} w_i + f(x_{it1}, \ldots, x_{itj}) + e_{it} \quad (4)$$

$$\begin{pmatrix} e_{i1} \\ \vdots \\ e_{iT_i} \end{pmatrix} \sim Normal\,(0, R_i) \quad (5)$$

and

$$w_i \sim Normal\,(0, D) \quad (6)$$

$The\ e_{it}$ are random errors that are independent and not associated with the random effects, w_i. R_i is a non-diagonal matrix that allows an autocorrelation structure within the errors for an individual. The RE-EM tree uses a tree structure to estimate f as well as the individual-specific random intercept w_i. Compared with a linear mixed-effects model (where $f = x\beta$), the RE-EM tree has more flexible assumptions, which admit that the functional form of f is normally unknown. The RE-EM tree can also better handle with missing values and overfitting issues. The estimation process of a RE-EM tree is shown as below [7]:

1. Initially set the estimated random effects, $\hat{w}_i$ to zero.
2. Run iterations through the steps a–c until the estimated random effects, $\hat{w}_i$, converge by considering change in the likelihood or restricted likelihood function being less than the tolerance value.
    a. Fit a regression tree to the data to predict the response variable using the attributes, $(x_{it1}, \ldots, x_{itj})$, for objects i = 1, ..., I at times t = 1, ..., T_i. The tree includes a set of indicator features, I ($x_{it} \in g_v$), where g_v ranges over all the terminal nodes in the tree.
    b. Estimate the linear mixed-effects model, $y_{it} = Q_{it} w_i + I\,(x_{it} \in g_v)\mu_v + e_{it}$ using the response variable and the attributes.
    c. Extract the estimated random effects $\hat{w}_i$ from the estimated linear mixed-effects model.
3. Replace the predicted values of the response variable at each terminal node of the tree in the step 2a with the population-level predicted mean response $\hat{y}_i$ from the linear mixed-effects model in step 2b.

Any tree algorithm can be applied to step 2a. Sela and Simonoff [7] implemented the CART tree algorithm based on the R package – rpart in the step 2a and developed the R package, REEMtree. The RE-EM tree algorithm maximizes the reduction in sum of squares when splitting a node. Maximum likelihood or restricted maximum likelihood (REML) can be used in step 2b. The

splitting process continues as long as the improvement in proportion of variability being accounted for by the tree (termed complexity parameter), which determines the optimal size of the tree. In the example of Sela and Simonoff [7], the value of complexity parameter (cp) was set at least 0.001, and the number of observations in the node was set at least 20. A 10-fold cross validation was applied to prune the tree once the initial tree was settled. The final split of the tree had the largest cp value and obtained the minimized validation error that was less than one standard error above the minimized value. The RE-EM tree allows for autocorrelation within individuals, which may yield more effective models comparing with no autocorrelation structure [7].

2.2.2. Mixed-Effects Random Forest

Hajjem et al. [14] expanded upon the CART algorithm [3] and introduced a mixed-effects regression tree (MERT) approach for handling clustered data with a continuous outcome. MERT utilizes the expectation-maximization (EM) algorithm to estimate the random components. Subsequently, a standard tree is applied to estimate the fixed effects after removing the random component. This approach enables the examination of non-linear relationships between the fixed components and response values.

To enhance prediction accuracy, Hajjem et al. [8] further developed a mixed-effects Random Forest (MERF), where a Random Forest replaces the regression tree. This advancement incorporates the benefits of ensemble learning to improve predictions in the presence of random effects. Additionally, Hajjem et al. [15] extended the MERT approach to handle non-Gaussian response variables, introducing a generalized mixed-effects regression tree (GMERT) that can address classification problems.

The MERF algorithm can be defined as follows:

$$y_i = f(A_i) + Z_i w_i + e_i \quad (7)$$

$$w_i \sim N(0, D), \quad e_i \sim N(0, R_i) \quad (8)$$

$$i = 1, \ldots, n_i, \quad (9)$$

where $y_i = [y_{i1}, \ldots, y_{in_i}]^T$ is the n_i 1 vector of responses for the n_i observations in the cluster i, $A_i = [A_{i1}, \ldots, A_{in_i}]^T$ is the matrix of fixed effects attributes, and $f(A_i)$ is estimated using Breiman's Random Forest [9]. $Z_i = [Z_{i1}, \ldots, Z_{in_i}]^T$ represents the n_i q matrix of random effects attributes for the cluster i, $w_i = (w_{i1}, \ldots, w_{in_i})^T$ is the $q \times 1$ matrix of random effects coefficients for the cluster i, and $e_i = (e_{i1}, \ldots, e_{in_i})^T$ is the n_i 1 vector of errors. D is the covariance matrix of w_i, while R_i is the covariance matrix of e_i. In the MERF algorithm, $Z_i w_i$ is assumed linear with the response variable, the random component $Z_i w_i \; and \; e_i$ is assumed to be independent and normally distributed. The covariance matrix of the response is assumed to be $V_i = \text{Cov}(y_i) = Z_i D Z_i^T + R_i$, and V = Cov(y) = diag($V_1$,…, V_n), where y = $[y_1^T, \ldots, y_n^T]^T$. Another assumption is the between-clusters are independent. Fitting the MERF allows us to predict new observations in the clusters considering the cluster-level random effects. The correlation is assumed to occur only via the between-cluster variations, where R_i is diagonal ($R_i = \sigma^2 I_n$, i = 1,…, n).

The overall steps of the MERF algorithm, as described in Hajjem et al. [8], can be outlined as follows:

1. Set r = 0 and the initial values for the parameters, which are $\widehat{w}_{i(0)} = 0$, $\widehat{\sigma}^2_{(0)} = 1$, $\widehat{D}_{(0)} = I_q$.

2. Set r = r + 1. Update the response corrected for the random effects $y^*_{i(r)}$, random forest of the fixed effects $\hat{f}(A_i)_{(r)}$, the random component $\widehat{w}_{i(r)}$:

(i) Set $y^*_{i(r)} = Z_i \widehat{w}_{i(r-1)}$, $i = 1,\ldots, n$.

(ii) Build a RF with $y^*_{ij(r)}$ as the response and a_{ij} as the corresponding training set of attributes, $i = 1,\ldots, n$, $j = 1,\ldots, n_i$. The bootstrap training samples are repeatedly drawn from the training set ($y^*_{ij(r)}$, a_{ij}).

(iii) Estimate $\hat{f}(A_i)_{(r)}$ using the out-of-bag prediction of the RF, that is, estimate each $\hat{f}(a_{ij})$ using the bootstrap samples to build the trees not containing observation a_{ij}.

(iv) Set $\widehat{w}_{i(r)} = \widehat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} (y_i - \hat{f}(A_i)_{(r)})$, $i = 1,\ldots, n$, where $\hat{V}_{i(r-1)} = Z_i \widehat{D}_{(r-1)} Z_i^T + \widehat{\sigma}^2_{(r-1)} I_n$, for $i = 1,\ldots, n$.

3. Update $\widehat{\sigma}^2_{(r)}$ and $\widehat{D}_{(r)}$ following

$$\widehat{\sigma}^2_{(r)} = \frac{1}{N} \sum_{i=1}^{n} \{\hat{e}^T_{i(r)}\, \hat{e}_{i(r)} + \widehat{\sigma}^2_{(r-1)} [n_i - \widehat{\sigma}^2_{(r-1)}\, tr(\hat{V}_{i(r-1)})]\}$$

$$\widehat{D}_{(r)} = \tfrac{1}{N} \textstyle\sum_{i=1}^{n} \{\widehat{w}^T_{i(r)}\, \widehat{w}_{i(r)} + [\widehat{D}_{(r-1)} - \widehat{D}_{(r-1)}\, Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \widehat{D}_{(r-1)}]\},$$

where $\hat{e}_{i(r)} = y_i - \hat{f}(A_i)_{(r)} - Z_i \widehat{w}_{i(r)}$.

4. Iterate the previous steps until convergence. Apply the generalized log-likelihood (GLL) criterion to confirm the convergence:

$$\text{GLL}(f, w_i | y) = \textstyle\sum_{i=1}^{n} \{[y_i - f(A_i) - Z_i w_i]^T R_i^{-1} [y_i - f(A_i) - Z_i w_i] + b_i^T D^{-1} w_i + \log|D| + \log|R_i|\}$$

When predicting a new observation j from known cluster i, we can use the population-averaged RF prediction $\hat{f}(A_{ij})$ and the random component $Z_i \widehat{w}_i$. If a new observation is from an unknown cluster not included in the sample, we use only the population-averaged RF prediction.

3. METHODS

3.1. Data

For this study, the PISA 2018 data set provided by the Organization for Economic Co-operation and Development (OECD) was utilized. The PISA 2018 survey aimed to assess the knowledge and skills of 15-year-old students in the areas of mathematics, reading, and science across 79 participating countries and regions. Additionally, 52 countries administered a questionnaire regarding students' familiarity with information and communications technologies (ICT). In this particular study, the focus was solely on the students' reading competencies (PV1READ) as the response variable.

After addressing missing values, two countries with varying numbers of observations were selected for analysis: Kazakhstan (n_1 = 10,040) and the United States (n_2 = 2,592). In this study, a total of 31 attributes were considered, encompassing ICT-related attributes, reading attributes,

and other relevant student information. Table 1 provides a list of these attributes along with brief descriptions.

Table 1. Attributes Information

Attribute Name	Description
PV1READ	Student reading performance score (WLE)
ICTHOME	ICT available at home
ICTSCH	ICT available at school
ICTRES	ICT resources (WLE)
INTICT	Student interest in ICT (WLE)
COMPICT	Perceived ICT competence (WLE)
AUTICT	Perceived autonomy related to ICT use (WLE)
SOCIAICT	ICT as a topic in social interaction (WLE)
ICTCLASS	Subject-related ICT use during lessons (WLE)
ICTOUTSIDE	Subject-related ICT use outside of lessons (WLE)
ENTUSE	ICT use for leisure outside of school (WLE)
HOMESCH	Use of ICT for schoolwork activities outside of school (WLE)
USESCH	Use of ICT at school in general (WLE)
PERFEED	Perceived Feedback from teachers (WLE)
EMOSUPS	Parental emotional support perceived by student (WLE)
LMINS	Learning time (minutes per week)
ESCS	Index of economic, social and cultural status (WLE)
UNDREM	Meta-cognition: understanding and remembering
METASUM	Meta-cognition: summarizing
METASPAM	Meta-cognition: assess credibility
HEDRES	Home educational resources (WLE)
STIMREAD	Teachers' stimulation of reading engagement perceived by student (WLE)
ADAPTIVITY	Adaptation of instruction (WLE)
TEACHINT	Perceived teacher's interest in teaching (WLE)
JOYREAD	Joy/Like reading (WLE)
SCREADCOMP	Self-concept of reading: Perception of competence (WLE)
SCREADDIFF	Self-concept of reading: Perception of difficulty (WLE)
PISADIFF	Perception of difficulty of the PISA test (WLE)
PERCOMP	Perception of competitiveness at school (WLE)
PERCOOP	Perception of cooperation at school (WLE)
ATTLNACT	Attitude towards school: learning activities (WLE)
BELONG	Subjective well-being: Sense of belonging to school (WLE)

It is worth noting that certain attributes in the PISA 2018 data set were derived using transformed weighted likelihood estimates (WLE) techniques [16].
The formula of transformation is as below:

$$W_t' = \frac{W_o - \bar{W}_{OECD}}{\sigma_{W_{OECD}}}$$

where W_t' is the final metric of the WLE scores after transformation, W_o is the original WLEs in logits, $\bar{W}_{OECD}$ is the mean score based on the equally weighted OECD country samples, and $\sigma_{W_{OECD}}$ is the standard deviation of the initial WLEs for the OECD samples.

The PISA 2018 applied plausible values for each student reading competency. Plausible values refer to a possible range of student competencies. Wu [17] noted that "instead of obtaining a point estimate for θ, a range of possible values for a student's θ, with an associated probability for each of these values, is estimated. Plausible values are random draws from this (estimated) distribution

for a student's θ. This distribution is referred to as the posterior distribution for a student's θ." (p. 116).

In this study, several attributes were selected that pertained to student engagement with teachers. These attributes encompassed aspects such as teachers' ability to stimulate reading engagement (STIMREAD), students' perception of teacher feedback (PERFREED), and students' perception of their teacher's interest in teaching (TEACHINT). Additionally, attributes related to students' meta-cognitive skills in reading were considered, including attributes such as understanding and remembering (UNDREM), summarizing (METASUM), assessing credibility (METASPAM), and enjoyment of reading (JOYREAD).

Other attributes related to learning included the amount of time spent on test language learning (LMINS), student adaptivity in test language lessons (ADAPTIVITY), and students' self-concept of reading, which encompassed their perception of competence (SCREADCOMP) and difficulty (SCREADDIFF). The study also took into account students' perception of the difficulty of the PISA 2018 test (PISADIFF).

Regarding students' background information, various attributes were analyzed. The index of student economic, social, and cultural status (ESCS) in the PISA 2018 data set was computed, taking into consideration factors such as parents' highest level of education, highest occupational status (HISEI), and home possessions (e.g., number of books). Other attributes included household possessions such as home educational resources (HEDRES) and parental emotional support (EMOSUPS).

To examine the impact of the school environment on student learning, attributes representing students' perceptions of the school were considered. These attributes encompassed students' perception of school competitiveness (PERCOMP), school cooperation (PERCOOP), attitude towards school (ATTLNACT), and the school climate as assessed by the scale measuring students' sense of belonging to school (BELONG).

3.2. Data Analysis

Two countries' data were extracted from the raw data set and treated as separate individual data sets. Prior to analysis, these data sets underwent a cleaning process to remove missing and noisy data points. Each data set was then divided into a 70% training set and a 30% testing set using random resampling without replacement within clusters. The training data sets were utilized to construct the RF regression, RE-EM tree, MERF, and HLM models. On the other hand, the testing data sets were not involved in the model development phase but were used to assess the performance of the models created during the training phase. In applying RF regression, RE-EM tree, MERF, and HLM, each clustered data set took into account the fixed effects of the selected attributes as well as the variability associated with the schools.

3.2.1. Building a RF model

The *randomForest* package [18] in R (version 3.5.2) was applied to implement the RF algorithm. The following hyperparameters of RF were applied in the tuning process:

1) Number of trees (*ntreeTry*). The default setting of number of trees (*ntreeTry* = 500) was adopted. In this study, 500 trees were sufficient to produce solid results.

2) The *stepFactor* is the value by which the number of features sampled when constructing each tree (mtry) is inflated or deflated. This value was set as 1.5.

3) The improvement value in the minimum out-of-bag (OOB) error (*improve*) to continue the search was set as 0.01.

4) Number of features sampled when constructing each tree (*mtry*). The default value of *mtry* was calculated using the formula, *mtry* = number of attributes / 3. The starting value of *mtry* follows *mtry* = default value / *stepFactor*. The ending value of *mtry* follows *mtry* = default value * *stepFactor*. Therefore, we used tuneRF function to confirm the best value of *mtry* based on the OOB error. In both the Kazakhstan and USA data sets, the tuning process showed that *mtry* = 7 was the optimal value.

3.2.2. Building a RE-EM Model

The *REEMtree* package [19] in R (version 3.5.2) was applied in the analyses. In the RE-EM tree analyses, 10-fold cross validation was applied when building the models, and complexity parameter (cp) was set as 0.01 for pruning the trees in order to select the optimal tree size based on the lowest cross validation error.

3.2.3. Building a MERF Model

The *merf* package in Python (version 3.8) was used to run the MERF regression. In this study, we set 300 trees generated in the random forest and 50 as the maximum number of iterations until convergence for both sampling data sets.

3.2.4. Applying HLM

The HLM method was conducted in R (version 3.5.2) using the package *lme4* [20]. The adjusted and conditional Intraclass Correlation Coefficient (ICC) was first run for each data set to estimate the variance explained by the school clustered structure. A random intercept model was employed for this study.

3.3. Evaluation Criteria

Once the RF regression, RE-EM tree, MERF, and HLM models were constructed, the testing data sets were employed to assess the performance of these models. Various evaluation metrics were utilized to measure the disparities between the predicted values and the actual values, including the mean square error (MSE), mean absolute error (MAE), mean absolute percent error (MAPE), and Accuracy (calculated as 100% minus MAPE). These metrics have been widely employed in previous research studies to evaluate model performance (e.g., [21]). Below are the formulas of MSE, MAE, and MAPE:

$$MSE = \frac{1}{n}\sum_{a=1}^{n}(y_q - \hat{y_q})^2$$

$$MAE = \frac{1}{n}\sum_{a=1}^{n}|y_q - \hat{y_q}|$$

$$MAPE = \frac{1}{n}\sum_{a=1}^{n}|\frac{y_q - \hat{y_q}}{y_q}|$$

where n is the sample size, $\overline{y_a}$ is the actual value, $\overline{\hat{y_a}}$is the predicted value. Smaller values of MSE, MAE, and MAPE indicate smaller discrepancies between the estimated model and the actual data, indicating better model performance.

4. RESULTS

Based on the findings, the baseline models revealed intraclass correlations of 0.387 for Kazakhstan and 0.15 for the United States. This indicates that 38.7% of the variation in student reading achievement in Kazakhstan can be attributed to school effects, while for the United States, the school effects account for 15% of the variation in student reading scores.

For the United States dataset, the random intercept model identified seven significant ICT-related attributes (HOMESCH, INTICT, AUTICT, SOIAICT, ICTCLASS, ICTHOME, and ICTSCH) and three significant teacher-related attributes (PERFEED, STIMREAD, and TEACHINT) that influenced student reading achievement. Significant impacts on student reading were also observed for student reading-related attributes (UNDREM, METASUM, METASPAM, SCREADCOMP, and JOYREAD), as well as other attributes such as EMOSUPS, HEDRES, ESCS, PISADIFF, PERCOOP, and BELONG. The overall HLM model achieved an accuracy of 88.22% for the United States.

In contrast, the HLM model for Kazakhstan yielded different significant attributes. Attributes such as ENTUSE, USESCH, COMPICT, and ICTRES significantly influenced student reading scores in Kazakhstan, while HOMESCH, AUTICT, and ICTSCH were found to be insignificant. Other significant attributes for predicting Kazakhstan students' reading performance included LMINS, ADAPTIVITY, and SCREADDIFF, which were not significant in the United States dataset. ESCS and BELONG were found to be insignificant for Kazakhstan students' reading performance. Overall, the HLM model for Kazakhstan achieved an accuracy of 89.8%.

Regarding the RF models, they explained 49.43% of the variance in the United States dataset and 53.17% of the variance in the Kazakhstan dataset. The top five important attributes in the RF model for the United States were METASPAM, PISADIFF, ESCS, JOYREAD, and METASUM. In the Kazakhstan dataset, the most important attributes were METASUM, UDREM, PISADIFF, METASPAM, and SCREADDIFF. The accuracy of the RF models was 92.61% for the United States and 93.72% for Kazakhstan.

Comparatively, the RE-EM tree models achieved lower accuracies, with 86.72% for the United States and 89.03% for Kazakhstan. The RE-EM tree structures, as shown in Figure 1 and Figure 2, were simpler for the United States dataset compared to the Kazakhstan dataset. METASPAM, PISADIFF, and METASUM were significant attributes contributing to the modeling structures for both datasets.

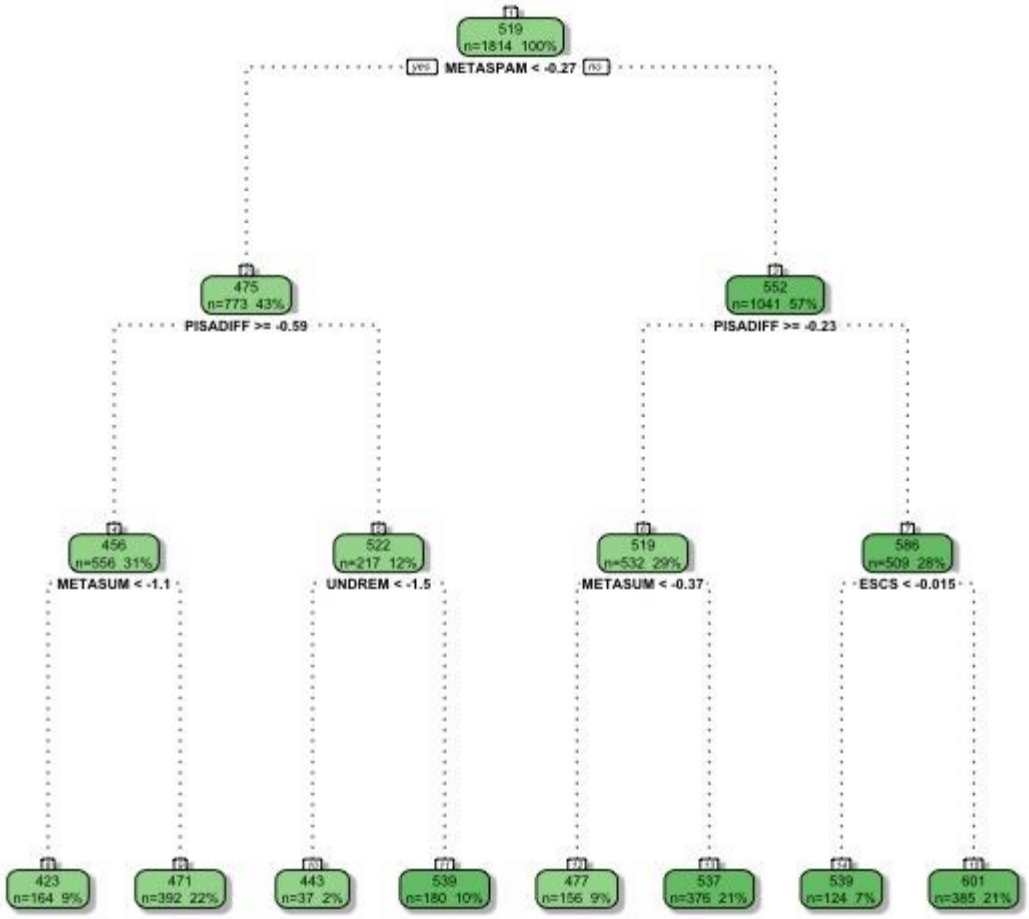

Figure 1. The United States RE-EM Tree Model Result. It shows the significant attributes and their thresholds. Those attributes are METASPAM, PISADIFF, METASUM, UNDREM, ESCS.

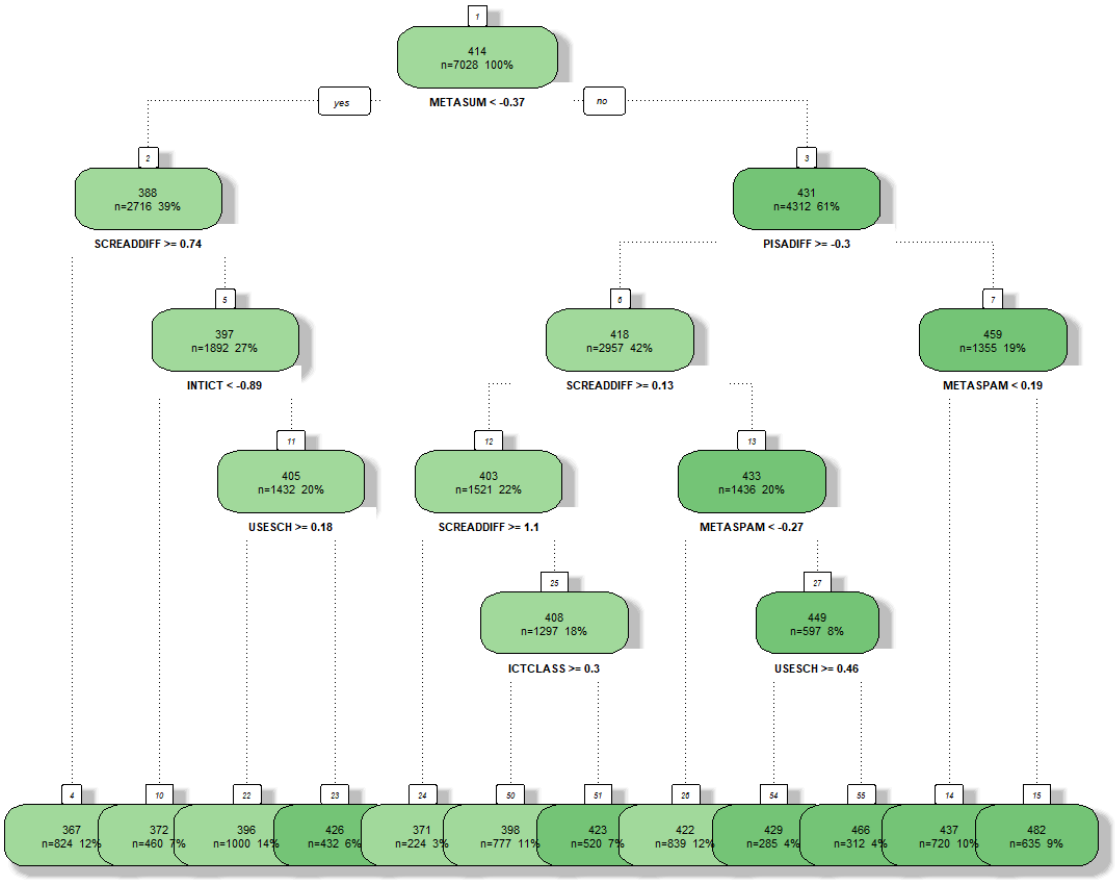

Figure 2. The Kazakhstan RE-EM Tree Model Result. It shows the significant attributes and their thresholds. Those attributes are METASUM, PISADIFF, SCREADDIFF, INTICT, USESCH, METASPAM, ICTCLASS.

The MERF models performed the best among the different methods for both datasets, achieving accuracies of 93.16% for the United States and 94.38% for Kazakhstan. Other evaluation metrics also indicated that the MERF models outperformed the other methods (see Table 2 and Table 3).

Table 2. The Evaluation Metrics Result of Each Model for the United States Data

	MSE	MAE	MAPE	ACCURACY
RF	2371.006	34.6963	0.0739	92.61%
RE-EM Tree	6238.66	62.8526	0.1328	86.72%
MERF	2207.5367	20.2245	0.0684	93.16%
HLM	4956.902	56.0686	0.1178	88.22%

Table 3**.** The Evaluation Metrics Result of Each Model for the Kazakhstan Data

	MSE	MAE	MAPE	ACCURACY
RF	1295.416	25.6777	0.0628	93.72%
RE-EM Tree	3227.529	45.0954	0.1097	89.03%
MERF	1143.1682	14.6682	0.0562	94.38%
HLM	2837.556	42.138	0.102	89.8%

Figures 3 and 4 further illustrated the influence of METASPAM, PISADIFF, and METASUM on students' reading performance, consistent with the results from the RF models. However, the MERF models slightly improved accuracy compared to the RF models in both datasets.

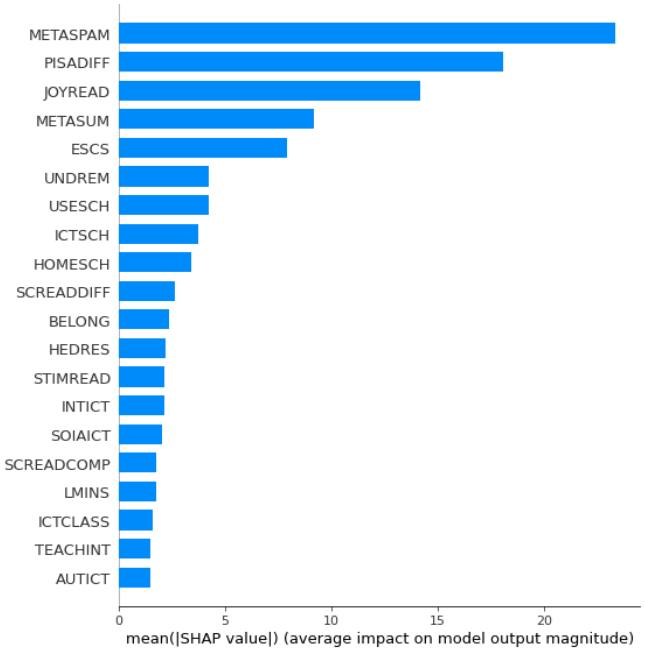

Figure 3. The Importance of Attributes in MERF Model for the United States Data.

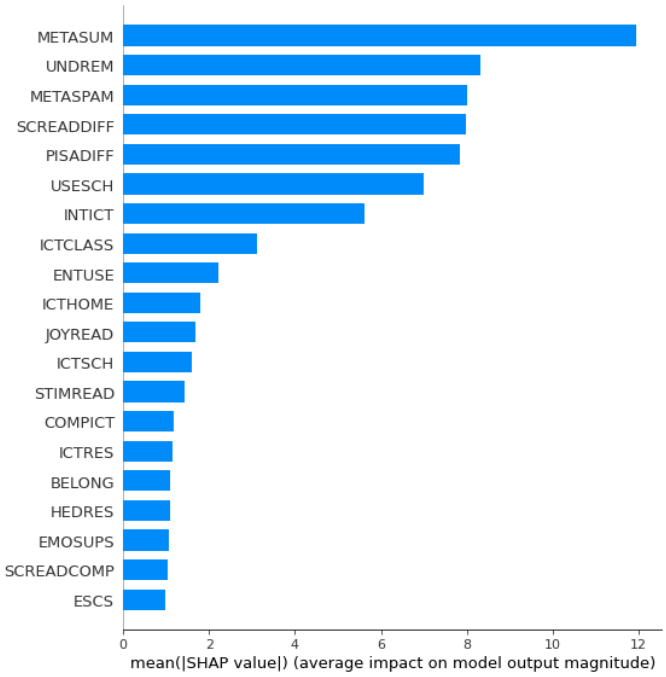

Figure 4. The Importance of Attributes in MERF Model for the Kazakhstan Data.

5. DISCUSSION

Among the methods applied, MERF proved to be the most accurate for both the United States and Kazakhstan datasets. MERF combines the advantages of the RF method, such as reducing overfitting, being less sensitive to outliers, easy parameter setting, and automatic variable importance generation. It is particularly suitable for clustering data as it considers both fixed and random effects of variables. The accurate predictions generated by MERF, using a bagging scheme, are valuable for predicting students' learning outcomes. A previous study by Pellagatti et al. [22] successfully applied a similar method called generalized mixed-effects Random Forest (GMERF) for predicting university student dropout.

However, MERF, like RF, has a major drawback in its "black box" nature, making it challenging to interpret the relationships between predictor and response variables. The ensemble tree structures hinder the interpretation of each tree, making it difficult to discern the exact directions and magnitudes of variables' impacts, although variable importance information is available. In this regard, the CART-based RE-EM tree method provides more interpretability of the results. RE-EM tree combines the advantages of both regression tree and linear mixed-effects regression algorithms. It is robust to outliers, as the tree-splitting process can isolate outliers in individual tree nodes. Additionally, RE-EM tree does not require preselected variables in high-dimensional datasets, providing flexibility in capturing data patterns. However, the method may generate unstable decision trees due to different splitting approaches adopted by the tree structure.

When comparing data mining methods with HLM in educational clustering data settings, data mining methods like MERF and RE-EM tree perform better for high-dimensional data, as they do not require specifying a functional form and can handle missing data values more effectively. The choice between MERF and RE-EM tree depends on the research study's objectives or applications. For instance, when developing an early alert system for identifying student dropouts or predicting course grades, MERF or GMERF can yield accurate predictions. These methods may also have great potential for use in other technologies in the future, such as intelligent tutoring systems, educational games, and recommender systems. On the other hand, when the main objective is to examine relationships among variables in big data for education, collected from technology systems or multiple sources, RE-EM tree may be more appropriate considering its interpretability.

Additionally, HLM remains a useful method for educational clustering data, especially when the data is not high-dimensional and does not have significant issues with outliers or missing values. For example, Xu et al. [23] applied HLM to investigate the relationship between students' ICT usage and learning performance in mathematics, science, and reading. Hew et al. [24] used HLM to predict student satisfaction with massive open online courses. Our study results demonstrated the advantage of applying HLM, which even showed slightly higher accuracy than the RE-EM tree model.

6. CONCLUSION

This study offers a comprehensive comparison of four statistical methods, namely RF, RE-EM tree, MERF, and HLM, in analyzing clustered educational data. The findings shed light on the strengths and limitations of each method and provide valuable guidance for researchers in the education field. Specifically, the study highlights the potential benefits of utilizing mixed-effects data mining methods like RE-EM tree and MERF to enhance model accuracy when dealing with clustered data structures. Researchers can leverage these insights to make informed decisions regarding the selection and application of statistical methods in their own studies.

One limitation of this study is its exclusive focus on educational data, specifically the PISA 2018 dataset. Future studies should expand their scope by testing these statistical methods on diverse datasets from other fields to validate the findings. Additionally, there is a need for further development of the algorithms to address their limitations in terms of interpretability. Improving the transparency and understanding of the models is crucial for their broader application and practical utility.

REFERENCES

[1] X. Hu, Y. Gong, C. Lai, & F. K. Leung, "The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis," *Computers & Education*, 125, 1-13, 2018.

[2] S. Park & W. Weng, "The relationship between ICT-related factors and student academic achievement and the moderating effect of country economic index across 39 countries," *Educational Technology & Society*, 23(3), 1-15, 2020.

[3] L. Breiman, J. H. Friedman, R. A .Olshen, & C. J. Stone, *Classification and Regression Trees*. Wadsworth and Brooks/Cole: Monterey, CA, USA, 1984.

[4] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, & J. Friedman, "Unsupervised learning," *The elements of statistical learning: Data mining, inference, and prediction*, 485-585, 2009.

[5] L. Breiman, "Bagging predictors," *Machine learning*, *24*(2), 123-140, 1996.

[6] A. Sandoval, C. Gonzalez, R. Alarcon, K. Pichara, & M. Montenegro, "Centralized student performance prediction in large courses based on low-cost variables in an institutional context," *The Internet and Higher Education*, 37, 76-89, 2018.

[7] R. J. Sela & J. S. Simonoff, "RE-EM trees: a data mining approach for longitudinal and clustered data," *Machine learning*, 86(2), 169-207, 2012.

[8] A. Hajjem, F. Bellavance, & D. Larocque, "Mixed-effects random forest for clustered data," *Journal of Statistical Computation and Simulation*, 84(6), 1313-1328, 2014.

[9] L. Breiman, "Random forests," *Machine learning*, *45*(1), 5-32, 2001.

[10] M. Fernández-Delgado, M. Mucientes, B. Vázquez-Barreiros, & M. Lama, "Learning analytics for the prediction of the educational objectives achievement," in *IEEE Frontiers in Education Conference (FIE) Proceedings*, Oct. 2014, pp. 1-4.

[11] X. Chen & H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, *99*(6), 323-329, 2012.

[12] J. R. Winitzky-Stephens, & J. Pickavance, "Open educational resources and student course outcomes: A multilevel analysis," *International Review of Research in Open and Distributed Learning*, 18(4), 35-49, 2017.

[13] H. Woltman, A. Feldstain, J. C. MacKay, & M. Rocchi, "An introduction to hierarchical linear modelling," *Tutorials in quantitative methods for psychology*, 8(1), 52-69, 2012.

[14] A. Hajjem, F. Bellavance, & D. Larocque, "Mixed effects regression trees for clustered data," *Statistics & probability letters*, 81(4), 451-459, 2011.

[15] A. Hajjem, D. Larocque, & F. Bellavance, "Generalized mixed effects regression trees," *Statistics & Probability Letters*, 126, 114-118, 2017.

[16] T. A. Warm. T. A., "Weighted likelihood estimation of ability in item response theory," *Psychometrika*, 54(3), 427-450, 1989.

[17] M. Wu, "The role of plausible values in large-scale surveys," *Studies in Educational Evaluation*, 31(2-3), 114-128, 2005.

[18] A. Liaw, & M. Wiener, "Classification and regression by randomForest," *R news*, 2(3), 18-22, 2002.

[19] R. J. Sela, J. S. Simonoff, & W. Jing, "Package "REEMtree": Regression Trees with Random Effects for Longitudinal (Panel) Data," *R Foundation for Statistical Computing: Vienna, Austria*, 2021.

[20] D. Bates, M. Maechler, & B. Bolker, "Walker., S. Fitting linear mixed-effects models using lme4," *J Stat Softw*, 67(1), 1-48, 2015.

[21] De Myttenaere, B. Golden, B. Le Grand, & F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, 192, 38-48, 2016.

[22] M. Pellagatti, C. Masci, F. Ieva, & A. M. Paganoni, A. M., "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, 2021.
[23] X. Hu, Y. Gong, C. Lai, & F. K. Leung, "The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis," *Computers & Education*, 125, 1-13, 2018.
[24] K. F. Hew, X. Hu, C. Qiao, & Y. Tang, "What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach," *Computers & Education*, 145, 103724, 2020.
[25] R. Jindal & M. D. Borah, "A survey on educational data mining and research trends," *International Journal of Database Management Systems*, 5(3), 53, 2013.
[26] A. Hershkovitz & R. Nachmias, "Online persistence in higher education web-supported courses," *The Internet and Higher Education*, 14(2), 98-106, 2011.
[27] R. Asif, A. Merceron. S. A. Ali, & N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, 113, 177-194, 2017.
[28] J. L. Hung, B. E. Shelton, J. Yang, & X. Du, "Improving predictive modeling for at-risk student identification: A multistage approach," *IEEE Transactions on Learning Technologies*, 12(2), 148-157, 2019
[29] W. Weng & W. Luo, "Exploring the influence of students' ICT use on mathematics and science moderated by school-related factors," *Journal of Computers in Mathematics and Science Teaching*, 41(2), 163-185, 2022.

AUTHORS

Wenting Weng is an instructional designer at Johns Hopkins University. She pursued her Ph.D. from Texas A&M University. Her research interests include educational data mining, learning analytics, and emerging educational technology, such as game-based learning and artificial intelligence in education.

Wen Luo is a Professor at the Department of Educational Psychology, Texas A&M University. Her research interests include growth modeling of longitudinal data, modeling of data with complex multilevel structures, and quantitative methods for teacher and program evaluations.