Behavior Research Methods (2023) 55:2109–2124
https://doi.org/10.3758/s13428-022-01910-8

Comparing the prediction performance of item response theory and machine learning methods on item responses for educational assessments

Jung Yeon Park[1,2] · Klest Dedja[3] · Konstantinos Pliakos[3] · Jinho Kim[2,4] · Sean Joo[2,5] · Frederik Cornillie[2] · Celine Vens[3] · Wim Van den Noortgate[2]

Accepted: 16 June 2022 / Published online: 11 July 2022

Abstract

To obtain more accurate and robust feedback information from the students' assessment outcomes and to communicate it to students and optimize teaching and learning strategies, educational researchers and practitioners must critically reflect on whether the existing methods of data analytics are capable of retrieving the information provided in the database. This study compared and contrasted the prediction performance of an item response theory method, particularly the use of an explanatory item response model (EIRM), and six supervised machine learning (ML) methods for predicting students' item responses in educational assessments, considering student- and item-related background information. Each of seven prediction methods was evaluated through cross-validation approaches under three prediction scenarios: (a) unrealized responses of new students to existing items, (b) unrealized responses of existing students to new items, and (c) missing responses of existing students to existing items. The results of a simulation study and two real-life assessment data examples showed that employing student- and item-related background information in addition to the item response data substantially increases the prediction accuracy for new students or items. We also found that the EIRM is as competitive as the best performing ML methods in predicting the student performance outcomes for the educational assessment datasets.

Keywords Item response theory · Explanatory item response model · Machine learning · Background information · Prediction performance · Educational assessment

✉ Jinho Kim
jinhokim@uos.ac.kr

Jung Yeon Park
jpark233@gmu.edu

Klest Dedja
klest.dedja@kuleuven.be

Konstantinos Pliakos
konstantinos.pliakos@kuleuven.be

Sean Joo
sjoo@ku.edu

Frederik Cornillie
frederik.cornillie@kuleuven.be

Celine Vens
celine.vens@kuleuven.be

Wim Van den Noortgate
wim.vandennoortgate@kuleuven.be

[1] College of Education and Human Development, George Mason University, 4400 University Dr, Fairfax, VA 22030, USA

[2] KU Leuven, Campus KULAK, Faculty of Psychology and Educational Sciences and itec, imec research group, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium

[3] KU Leuven, Campus KULAK, Department of Public Health and Primary Care and itec, imec research group, Etienne Sabbelaan 53, 8500 Kortrijk, Belgium

[4] Graduate School of Education and Urban Bigdata•AI Institute, University of Seoul, 163 Seoulsiripdaero, Dongdaemun-gu, Seoul 02504, South Korea

[5] Department of Educational Psychology, University of Kansas, 1450 Jayhawk Blvd, Lawrence, KS 66045, USA

Introduction

Educational assessment is the systematic process of evaluating students' knowledge, skills, and abilities to find better ways to refine teaching and learning. In practice, however, educational environments have been shown to vary across schools, classes, and course delivery modes (e.g., emergency remote context due to the COVID-19), making it difficult to create assessments that incorporate the nuances of instructional content in the test items (Jiao & Lissitz, 2020). Furthermore, an educational assessment that does not consider the needs of diverse student groups provides educators and administrators with limited feedback on their educational design decisions. That is to say, the context-aware assessment is necessary to evaluate and predict students' learning outcomes more accurately and figure out strategies to refine and advance educational practices. Examples of contextual information that can be considered in the assessment include students' demographic characteristics (e.g., gender, age, and primary language), their prior knowledge level (e.g., previous courses taken), and components of the test design (e.g., item format and cognitive domains) they interact with.

In that regard, large-scale assessments of student learning (e.g., Trends in International Mathematics and Science Study, the Programme for International Student Assessment) have been considered to provide a window to the domain-specific knowledge and generate information about students' achievements in relation to some of the correlates of learning, such as student background, attitude, and perceptions, and perhaps school and home characteristics (Anderson et al., 2007). While the primary source of data for the student assessments is the information obtained from student responses to a set of test items, a rich source of data that is often neglected when analyzing the assessment data is the variety of background information related to test-takers, test designs, educators, and schools. Utilizing such information in addition to the student responses on a test helps to understand and predict students' performance outcomes in the educational assessment and hence to optimize teaching and learning strategies in educational practices.

To obtain more accurate and robust feedback information from the students' assessment outcomes and to communicate it to students, educational researchers and practitioners must critically reflect on whether the existing methods of data analytics are capable of retrieving the information provided in the database. This study pays attention to both theory-based and data-driven methods to investigate the prediction performance: one is an item response theory (IRT) method, and the others are machine learning (ML) methods. IRT is a theory-based psychometric approach to analyze categorical item response data typically obtained from educational assessments. In basic IRT models, such as the Rasch model, the probability of a correct response is modeled as a nonlinear function of students' latent abilities and items' difficulty parameters. In the realm of IRT, explanatory item response models (EIRM; De Boeck & Wilson, 2004) aim to explain and predict the parameters at either the student side, the item side, or both sides of the item response data, by incorporating student- and item-related properties or features (i.e., background information) as explanatory variables in the statistical model. Although the EIRM was originally developed to enhance explanatory inferences from the data, it can be used for predictive purposes in that explanation and prediction are inherently conflated in a statistical model (Shmueli, 2010). For instance, an extended version of the EIRM was used to predict dichotomous and/or polytomous item difficulties for the newly developed items (e.g., Kim & Wilson, 2020). Also, in the e-learning assessment, the EIRM was used to alleviate the cold-start problem in prediction that occurs when a new student joins an adaptive e-learning environment that aims to meet the student's learning needs through adaptive item selection (e.g., Park et al., 2019). Using the EIRM with background information, the parameters of item difficulties and/or students' latent abilities are predicted and thereby the categorical item responses are predicted based on the probability determined from the parameter estimates. Provided that the item responses are dichotomous, the predicted values are equal to 1 (correct answer) or 0 (incorrect answer), which implies that the EIRM can do classification to predict a binary class of the item responses.

ML is a modern data-driven approach to develop computationally efficient and accurate predictive algorithms (Shmueli, 2010). Regarding the item response prediction, there has been a substantial increase in exploring the potential of ML methods. Among the ML families, supervised learning (Horvitz & Mulligan, 2015) uses an available data set in order to obtain a model where the corresponding learning process is referred to as training. In the context of educational assessment, the training set includes the data generated through learner–item interactions that are described by students (e.g., gender) and items (e.g., item difficulty); and the labels refer to the student–item interaction. Using this training data, one can build a function (model) which performs target predictions (output variable) for new observations (i.e., student responses to items unsolved; Witten et al., 2011). The most common prediction tasks include classification (predicting categorical values) and regression (predicting numerical values) for the new observations.

Previously, several studies have applied ML methods to the contexts of educational assessment and most of them

employed the effective ML methods to develop predictive models for students' performance outcomes, mostly binary item responses. These models are often trained over student- and/or item-related background information (or features). The task is often to perform student grades or dropout predictions. For example, Kotsiantis (2012) showed that a decision-support system can be built to predict students' performance outcomes. More specifically, the system was trained on students' demographic information and marks in written assignments, addressing student grade prediction as a regression problem. In addition, in the study by Rovira et al. (2017), ML was employed for students' grades and dropout intention prediction. The authors proposed a personalized course recommendation model based on the data from computer science, law, and mathematics courses and investigated course preferences as well as course completeness ratios using decision tree learning (Hsia et al., 2008). Lykourentzou et al. (2009) proposed a dropout prediction method for e-learning courses using a combination of multiple ML techniques.

Furthermore, recent studies have attempted to create methodological connections between IRT and ML methods. For example, Bergner et al. (2012) derived that (multidimensional) IRT models can be viewed as a specific instance of collaborative filtering algorithms. Pliakos et al. (2019) proposed a hybrid approach that combines person ability and item difficulty estimates from IRT into ML methods using student- and item-related information to improve the accuracy of item response prediction. Gonzalez (2020) compared IRT and ML approaches for diagnostic assessment as well as individual classification and concluded that ML methods using logistic regression and random forest could have comparable classification accuracy to the psychometric methods using estimated IRT scores.

Despite the increasing number of studies in the topic, to our knowledge, there are relatively few studies that have compared and contrasted IRT and ML methods considering student- and item-related background information to predict student outcomes for educational assessments. Furthermore, little is known about prediction performance of both methods to examine potential prediction scenarios in educational assessments. Given that predicting student outcomes in a test is forecasting unrealized or missing item responses, one may be interested in predicting (a) unrealized responses of new students to existing items where there are no historical data about their performance (new student cold-start); (b) unrealized responses of existing students to new items that haven't been attempted by anyone in the assessment system (new item cold-start); and (c) missing responses of existing students to the items that already exist in the system.

In this paper, we approach these prediction scenarios using a range of supervised learning methods—decision tree learning, similarity-based methods, tree-ensemble learning, statistical classifier, and neural networks—as well as an EIRM to predict a binary class (correct or incorrect) of students' responses to item-based assessments. Each prediction method is evaluated through cross-validation approaches under the three (above-mentioned) prediction scenarios. In a simulation study, we further examine factors that affect their prediction performance in various data conditions. Next, we demonstrate their application by means of two educational assessment datasets in real-life settings. We end with conclusions and a discussion.

Prediction methods

Item response theory (IRT) method

As explanatory IRT (EIRT) modeling, the EIRM enables explanatory and predictive inferences from assessment data by incorporating student- and/or item-related background information (i.e., features or properties) as explanatory variables in the statistical model. Compared to descriptive IRT models such as a Rasch model which simply describes (differences in) student abilities and item difficulties, the EIRM approach implies the use of person explanatory, item explanatory, and doubly explanatory IRT models (De Boeck & Wilson, 2004), which can explain differences at the student side, item side, and both sides of the item response data, respectively. Once the effects of the explanatory variables are estimated from the assessment data through a relevant EIRM, one can use the estimates to predict person parameters (student latent abilities or proficiencies) and/or item parameters (item difficulties). These predicted parameters can be used in turn to compute the item response probabilities from which derive students' assessment outcomes via a stochastic process, and also the categorical item responses are predicted from the computed probabilities inversely.

Given the dichotomous (binary) item responses in the real-life assessment data examples and the three prediction scenarios we have considered, we focus on a doubly explanatory dichotomous IRT model (see De Boeck & Wilson, 2004). In addition to random person effects, the model includes random item effects, taking into account that in practice there is typically no perfect explanation/prediction of students' abilities and items' difficulties based on observable background information (De Boeck, 2008). This model is regarded as a crossed random effects model, namely cross-classification multilevel logistic model (Van den Noortgate et al., 2003). Because item responses (i.e., first-level observations) are nested in each of both persons and items (i.e., second-level units) but these two are not nested within each other, allowing for random effects on both parameters in the model makes them crossed; the two random effects on persons and items from the item responses are cross-classified.

Table 1 Presentation of the tuned parameters related to each method

Method family	Method	Hyperparameters
Item response theory	Explanatory item response model (EIRM)	Not applicable
Decision tree learning	Decision tree (DT)	Minimum samples per leaf {5,25,50,75,100}
Tree ensemble learning	Random forest (RF)	Min samples per leaf {1, 2, 5}, # trees: 200
Tree ensemble learning	Gradient boosting (GB)	Max tree depth {3, 6}; learning rate {0.001, 0.01, 0.1}; number of estimators {100, 200}
Similarity-based method	k-Nearest neighbors (k-NN)	Number of neighbors {5,10,25,50,75,100}
Statistical classifier	Quadratic discriminant analysis (QDA)	Not applicable
Neural Network	Multi layer perceptron (MLP) classifier	# hidden layers {2, 3}, neurons per layer {10, 20, 25, 40, 50}; learning parameter α (L2 regularization term) {0.00001, 0.0001, 0.001, 0.01, 0.1}

Thus, this extended doubly explanatory dichotomous IRT model is a latent regression linear logistic test model with random item errors, which can predict both student proficiencies and/or item difficulties and predict in turn student outcomes by employing student- and/or item-related information. This model will be hereafter referred to as the EIRM or the EIRT, for the purpose of calling it simply in this paper. A mathematical expression of the EIRM we used as an IRT method here is formulated as follows:

$$\ln\frac{P(y_{pi}=1)}{P(y_{pi}=0)} = Logit\,P(y_{pi}=1) = \alpha_0 + \left(\sum_{j=1}^{J}\omega_j z_{pj} + \epsilon_p\right) - \left(\sum_{k=1}^{K}\gamma_k x_{ik} + \epsilon_i\right), \quad (1)$$

where y_{pi} is the dichotomous item response of student p ($p=1, \ldots, P$) on item i ($i=1, \ldots, I$), α_0 is the overall intercept representing the overall logit of the probability of a correct response over students and items (when all background information variables are equal to zero), ω_j is the regression weight or the effect of student-related background information variable j on student proficiencies, z_{pj} is the value of student p on student-related background information variable j ($j=1, \ldots, J$), ϵ_p is a random noise/error or residual on student proficiencies, $\epsilon_p \sim N\left(0, \sigma_p^2\right)$, γ_k is the regression weight or the effect of item-related background information variable k on item difficulties, x_{ik} is the value of item-related background information variable k ($k=1, \ldots, K$) for item i, and ϵ_i is an random noisy/error on item difficulties, $\epsilon_i \sim N\left(0, \sigma_i^2\right)$. We used the R package, "lme4" (Bates et al., 2014) to fit the EIRM to simulated data and two real-life assessment datasets.

Machine learning (ML) methods

In the machine learning set-up, we treated item- and student-related background information as input and the student response y as the binary output to be predicted. In this study, we explored a variety of ML algorithms that are extensively employed for classification tasks (See Table 1 below). We chose the following algorithms not only because of their popularity but also with the aim to provide a diverse enough comparison pool of different ML methods. These methods are well established in the field of machine learning, representing prominent families of ML models, such as neural networks, decision tree learning, tree-ensemble learning, similarity-based methods, and statistical methods. It is worth noting that because of the "no free lunch" theorem (Wolpert & Macready, 1997), we cannot know in advance which algorithm will perform best on a given set of data, and we are therefore encouraged testing multiple models.

The first method we considered is decision tree learning (DT; Breiman et al., 2017; Quinlan, 1986). Here, the learning process was achieved by building a decision tree, a flowchart-like structure composed of *nodes* and *edges* which connect them, as shown in Fig. 1. The initial node is called root node and it contains all the training samples, here students and items. From the root of the tree, every node is recursively split based on a splitting criterion until final nodes (*leaves*, without an output edge) are reached. The labels corresponding to samples within each leaf are then used to determine the predicted label for (future) samples that end up in the same leaf at the end of the partitioning procedure. The most common labeling procedure follows a majority rule approach: the most common label of the leaf is used to predict labels for new samples within the same leaf.

DTs are popular due to their scalability and interpretability advantages. However, they often suffer from instability in their predictions and from overfitting. Although decision trees are considered relatively weak classifiers, when combined with ensemble learning they can provide state of the art results (Fernández-Delgado et al., 2014). These ensemble methods build many decision trees and the responses from such trees are combined to get the final output of the model. The trees therefore contribute all to the final prediction for a new sample according to rules determined by the ensemble method. In this study, we decided to include

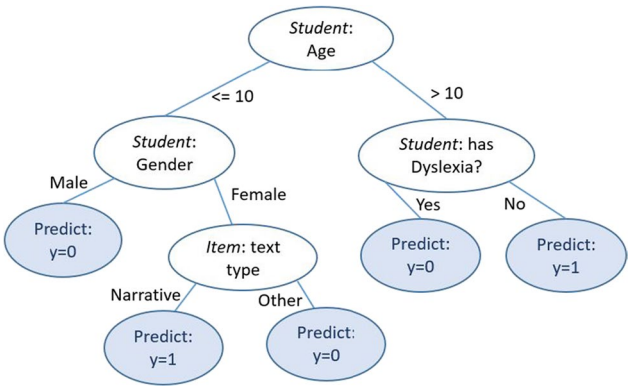


Fig. 1 A simple illustration example of a decision tree

two tree-ensemble methods: Random Forest (RF; Breiman, 2001) and Gradient Boosting (GB; Chen & Guestrin, 2016; Friedman, 2001).

An important characteristic of the RF method is the diversity that is enforced among the trees. This is obtained by using bootstrap replicates of the training set and random selection of the features (or background information) describing the samples. More specifically, each decision tree of the ensemble is constructed on a random subset of the training set. Moreover, every node of that tree is split by computing the best possible split among a random subset of selected feature candidates, leading to further diversification. The final prediction is yielded as the average of the predictions of individual trees. Tree-based learning has many advantages, such as scalability and computational efficiency. GB is another ensemble model based on trees. In this method, trees are built in succession: the first tree represents an initial coarse fit and every subsequent tree represents a fit of the prediction error made until the previous step. The procedure continues until a large number of trees is built, and the final prediction of the model is a (weighted) sum of the single-tree predictions. Gradient boosting trees and especially a variant denoted as eXtreme Gradient Boosting (XGBoost), are widely utilized and respected by the ML community[1].

Apart from the tree-ensembles methods, we take into consideration additional widespread classification algorithms. The k-nearest neighbors' classifier (k-NN; Altman, 1992) classifies new samples based on the most common class among their k nearest neighbors in the input (features or background information) space. Similarity is computed upon the values of the input features, such as student-related features (age, primary language) and item-related features (type of task, vocabulary). As a consequence, samples deemed to be similar are labeled as part of the same class.

[1] "XGBoost – ML winning solutions (incomplete list)" in https://github.com/dmlc/xgboost/tree/master/demo#machine-learning-challenge-winning-solutions

Another widespread classification algorithm is the quadratic discriminant analysis (QDA; Tharwat, 2016). This statistical classification technique considers a set of observations and groups them in classes with the same outcome following a quadratic decision surface. For each new observation, the QDA method calculates the probability of belonging to each class and assigns the label to the class with the highest probability. That is, assuming we are in a case where sample labels y_i either have the value "0" or "1", QDA assigns observations to the class "1" if:

$$P\left(y_i = {}^{\varepsilon}1^{\varepsilon}\right) \geq 0.5$$

and to the class "0" otherwise.

Finally, we considered one algorithm from the (Deep) Neural Network family. In particular, we employed the multi-layer perceptron (MLP; Hastie et al., 2009; Van Der Malsburg, 1986) classifier, as illustrated in Fig. 2, a feed-forward neural network with many possible configurations in its architecture (number of neurons, number of layers) and signal propagation (activation function, backpropagation) that can be chosen or tuned. In our example, we provide student- and item-related background information to the input layer, and the output layer predicts the probability of being part of class "0" or class "1". We opted for a network with a rectified linear unit (RELU) activation function and a stochastic gradient-based algorithm for weight optimization called "Adam" (Kingma & Ba, 2017).

Comparison procedure

Prediction scenarios

The overarching goal of this paper is to investigate the performance of IRT and ML methods in predicting a binary (correct or incorrect) class of students' item responses on the educational assessments. A total of seven prediction methods were compared through a simulation study (Sect. 4) and two real-life data examples obtained from university- and national- level summative assessments (Sect. 5). In the item response data from the educational assessments, unobserved response values are supposed to be predicted. The unobserved responses in the data are either unrealized or missing in the assessment system. Since responses of new students to new items have nothing to do with the assessment system, three prediction scenarios are considered for each item response dataset, as described in the three panels of Fig. 3:

- Predicting new students' (P_{new}) unrealized responses for existing items (I), we refer to this set-up as *(new) student scenario* from now on, illustrated in the left panel;

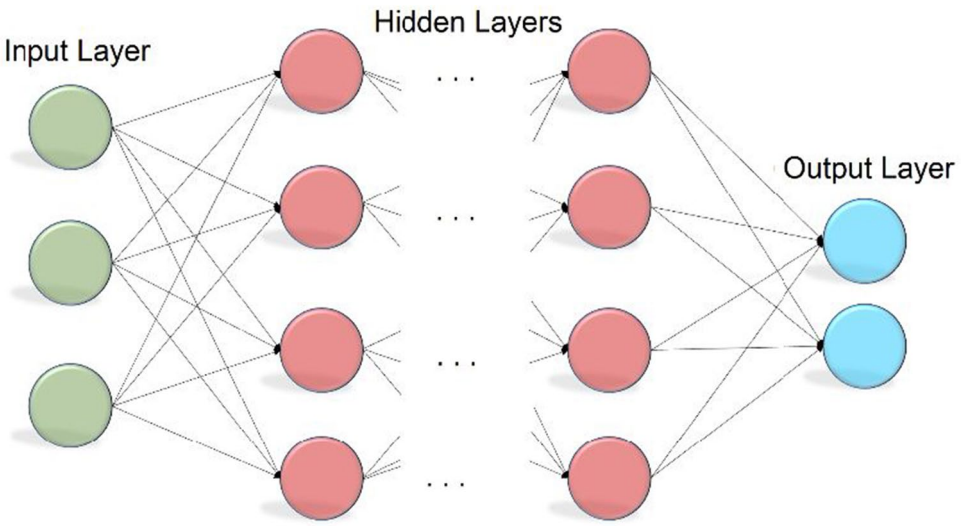


Fig. 2 A schematic illustration of a multi-layer perceptron with input, hidden and output layers

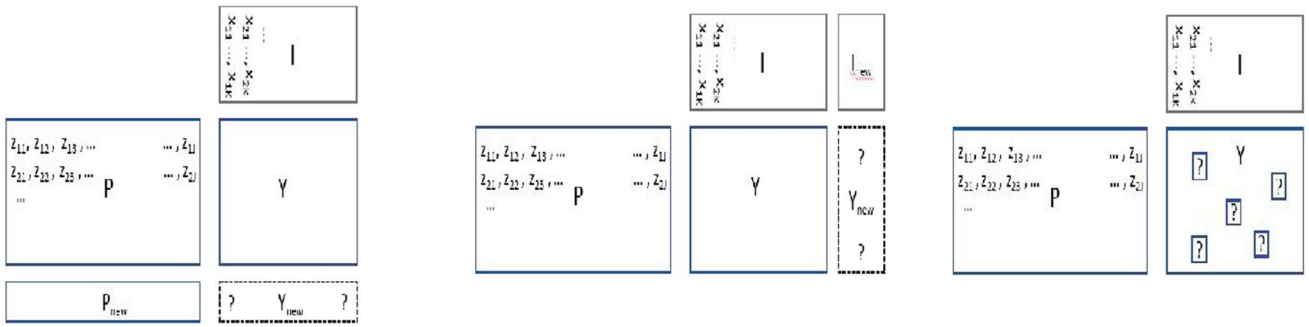


Fig. 3 Illustrations of the three prediction scenarios

- Predicting existing students' (P) unrealized responses for new items (I_{new}), the *(new) item scenario,* illustrated in the central panel;
- Predicting existing students' (P) missing responses for existing items (I), *student–item pair scenario,* illustrated in the right panel.

With regard to the ML methods, we addressed the data settings above as single-output (univariate) classification tasks. In order to achieve this, we constructed the data matrix as the Cartesian product of student and item samples. Each sample in our task is therefore a pair of a student (P) and an item (I). The data matrix is composed of |P| x |I| pair-samples and each pair is described by a concatenation of student-related and item-related background information. The construction of the data matrix is illustrated in Fig. 4.

Evaluation metrics

We considered three evaluation criteria including area under receiver operating characteristic (AUROC) curve, area under precision recall (AUPR) curve, and mean squared error (MSE). Note that the ROC curve represents the ratio between true-positive (TP) rate, $\left(\frac{TP}{TP+FN}\right)$and false-positive (FP) rate, $\left(\frac{FP}{FP+TN}\right)$ at various probability thresholds, where FN and TN indicate the number of false negatives and true negatives, respectively. The precision recall curve is defined as the precision, $\left(\frac{TP}{TP+FP}\right)$ against the recall, $\left(\frac{TP}{TP+FN}\right)$, again for various thresholds. In case of totally random predictions the AUROC value is approximately equal to 0.5 and AUPR is equal to the frequency of the positive class. For both measures, 1 is the value achievable by a model with perfect predictions. In

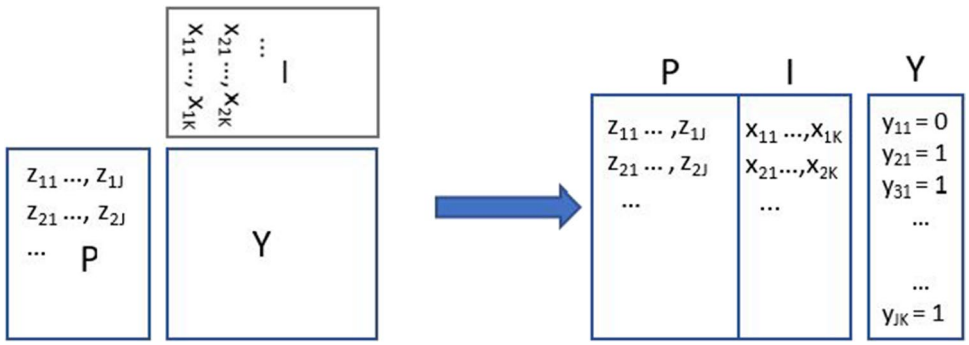


Fig. 4 An illustration of the data matrix construction for the ML methods

addition, the MSE is defined as $MSE(y_{pi}, \hat{y}_{pi}) = E(y_{pi} - \hat{y}_{pi})^2$, where y_{pi} is the observed response and $\hat{y}_{pi}$ is the predicted one.

Experimental protocol

We validated the performance of the prediction methods in a nested k-fold cross validation (CV) procedure with five inner folds for parameter tuning and ten outer folds for performance evaluation. This nested CV, albeit computationally expensive, is a good practice to avoid optimistic estimation of model performance and therefore reduce selection bias (Cawley & Talbot, 2010). We performed parameter tuning with grid search to increase performance of the algorithms; more details about the corresponding parameters and their tested values are found in Table 1.

Lastly, in order to test for statistically significant differences among all the methods, we followed the procedure suggested by Demšar (2006). In particular, we conducted a Friedman test (Friedman, 1940), based on the average ranks of each method's performance across 4 datasets × 3 prediction scenarios. If the omnibus test shows that there was a statistically significant difference at significance level .05 among the competing methods ($p \leq .05$), we further conducted a Nemenyi test (Nemenyi, 1963) for a post hoc comparison. The post hoc test computes a "Critical Difference" (CD; Demšar, 2006), also referred to as Critical Distance, threshold for a given significance level (again, a significance level of .05 was used), and if the difference between the average ranks of two methods is greater than the CD, the performance of the two is concluded to be statistically significantly different.

Simulation study

Design

To examine how predictive capability of the EIRM and ML methods are affected by different aspects of the assessment data, we conducted a cross-validation with simulated datasets. Table 2 shows four simulated datasets differentiated by the specific conditions of data size and degree of noise. To generate student p's response to item i, we used an EIRM; specifically, the student ability parameter (ϵ_{pm}) was generated to have multidimensionality (while the EIRM assumes unidimensionality for analyzing the data):

$$Logit\, P(Y_{pi} = 1) = \alpha_0 + \sum_{j=1}^{J} \omega_j z_{pj} + \boldsymbol{q}_i' \boldsymbol{\epsilon}_p - \left(\sum_{k=1}^{K} \gamma_k x_{ik} + \epsilon_i \right) \quad (4)$$

- ***Student component***. For student fixed effects, each dataset has $J = 15$ student-related variables, z_{pj} ($j = 1, \ldots, J$); and the corresponding coefficients ω_j were randomly sampled from independent univariate normal distributions, $N(0.2, 1)$. The intercept, α_0 was set at 1.2. For student random effects, true values for student p in M-dimensional ability space, $\boldsymbol{\epsilon}_p = (\epsilon_{p1}, \ldots, \epsilon_{pM})'$ is a vector of M student-specific deviations that were randomly sampled from $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (0, \ldots, 0)'$ and $\rho_{mm'} = 0.3$ ($m \neq m'$). And $\boldsymbol{q}_i = (q_{i1}, \ldots, q_{iM})'$ is a vector of M coefficients for item i that specify the relations between the item and each individual ability; each of them was randomly sampled from Bernoulli (.5).
- ***Item component***. For item effects, each dataset has K=10 item-related variables, x_{ik} ($k = 1, \ldots, K$); and the corresponding fixed coefficients γ_k were randomly sampled from $N(0.5, 1)$. For random item effects, true values for item i, ϵ_i were randomly drawn from $N(0.5, 1)$.

Based on the data-generation scheme, we want to examine the effects of three factors on their prediction accuracy and consistency, including (a) data size, (b) degree of noise in the student- and item-related variables. More specifically, we are interested in observing the effects caused by a reduction of the data size as well

Table 2 Summary of the four simulated datasets

Dataset description	Data size	Complexity
Small, 10%	Small (100 students–10 items)	Noise: 10%
Normal, 10%	Typical (Normal) (1000 students–100 items)	
Normal, 30%		Noise: 30%
Normal, 60%		Noise: 60%

In each dataset, tenfold CV for three prediction scenarios (new students, new items, and student–item pair scenarios) were conducted

as the effects of an increase of the degree of noise. For this purpose:

- ***Data size***. To examine the extent to which the prediction methods are affected by the shortage of data for training, we consider two types of data sizes in large-scale educational assessments: a typical (normal) size (1000 students – 100 items) and a small size (100 students – ten items). In comparing the two data sizes, we look at the two datasets with random noise level set at a *low* level (10%); and dimensionality in student ability set at a *moderate* level, so $M = 9$.
- ***Noise***. To examine the extent to which the prediction methods are affected by the different level of noise for training, we considered three levels of noise: 10% (low), 30% (moderate), and 60% (high). We define noise by the percentage of the random effect variance (ϵ_{pm}) compared to the total variance in the student component for each dimension, $\left(\sum_{j=1}^{J} \omega_j z_{pj} + \epsilon_{pm}\right)$, indicating the portion that is not explained by the set of student-related variables. The number of student-related variables was fixed at 15; similarly, the number of item-related background information variables was fixed at 10 as generally there is more information about students as compared to items.

Because the data-generation was carried out using the EIRM, one can expect that the EIRM may be more beneficial when comparing the performance of different prediction methods among IRT and ML approaches. To alleviate such potential problems, the student ability parameters were assumed to be multidimensional in the data generation process, whereas a unidimensional EIRM model was used in the analysis.

Results

Figures 5, 6 and 7 visualize results of response prediction of the seven prediction methods (one IRT and six ML methods) across the simulated datasets. Each figure includes three panels for the experimental scenarios introduced in Sect. 3.1: new student scenario (top), new item scenario (middle), and student–item pair scenario (bottom). Each figure consists of a set of bar charts for the performance metrics– AUPR (Fig. 5), AUROC (Fig. 6), and MSE (Fig. 7) (on the *y*-axis) for a combination of different competing methods and the simulation conditions (on the *x*-axis). Note that greater values of AUPR and AUROC and smaller values in MSE indicate better performance (i.e., predictive capability).

In general, we found that the prediction accuracy differs by the three prediction scenarios. Specifically, the best overall performance is seen under the new student–item pair scenario across methods (on average, AUPR = .861; AUROC = .9; MSE = 0.143), followed by the new item scenario (on average, AUPR = .827; AUROC = .872; MSE = 0.169) and the new student scenarios (on average, AUPR = .782; AUROC = .840; MSE = 0.185). Also, we found that EIRM shows the highest AUPR and AUROC and the lowest MSE across datasets (on average, AUPR = .886; AUROC = .918; MSE = 0.137). Among the ML methods, GB (on average, AUPR = .832; AUROC = .880; MSE = 0.157) and RF (on average, AUPR = .829; AUROC = .878; MSE = 0.158) show the best performance followed by MLP. The DT and QDA yield the worst performance.

A comparison of the small- and typical-sized datasets (with 10% noise and nine dimensions) suggests that performance of any method gets worse for small-sized dataset as compared to typical -sized dataset. In particular, DT seems to be vulnerable to the small-sized dataset with AUPR values less than or equal to .8 in the new student and item scenarios.

For the effect of random noise on performance, results highlight that the prediction accuracy drops with increasing levels of random noise. We observed a minimal difference between the setups with 30% noise compared to the ones with 10% noise and the difference increased when increasing the noise parameter to 60%. We also found that with a 60% noise level for new student scenario that reveals overall the lowest accuracy, EIRM still has AUPR values greater than .75, while the values are less than .70 for all MLs in the scenario; it suggests that the performance of the ML methods is more affected by the increasing random noise (weaker explanatory power of the student- and item-related background information) than EIRM, implying the robustness of EIRM.

The results of the statistical analysis are summarized in Fig. 8, where the test results in regard of AUROCs are visualized using the R package "scmamp" (Calvo & Santafé Rodrigo, 2016). In the figure, the average ranks of the methods are indicated by vertical lines (e.g., the average rank of EIRM is 1); in addition, the methods that are not statistically significantly different are connected by thicker horizontal segments. We found that EIRM was the best

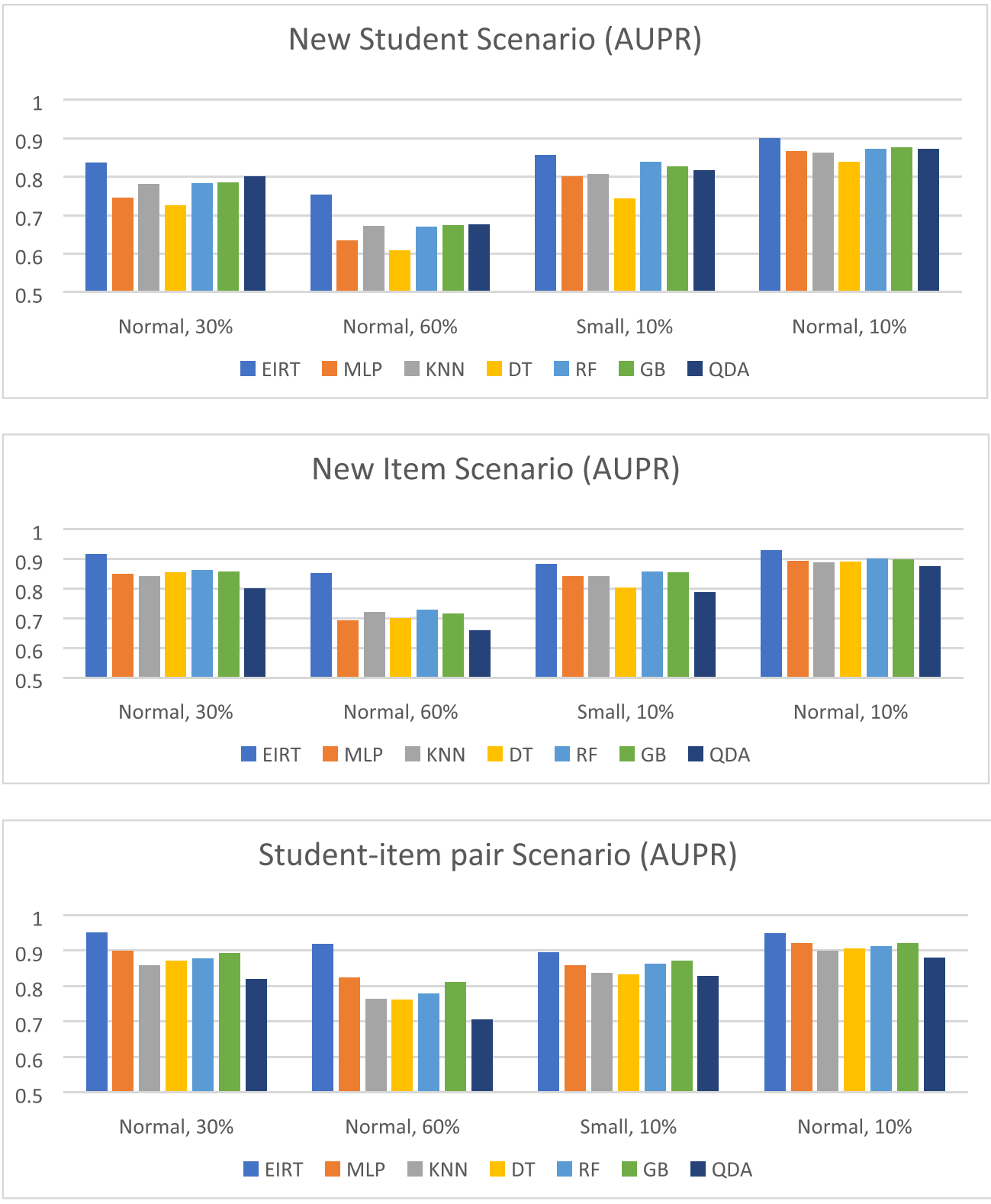


Fig. 5 Summary of simulation study: AUPR

Fig. 6 Summary of simulation study: AUROC

Fig. 7 Summary of simulation study: MSE

performing method in terms of the robustness and accuracy (ranked first), but it is not statistically significantly different from GB and RF. It is also confirmed that RF and MLP lied somewhere in the middle of the performance spectrum (with RF outperforming MLP, although not significantly). On the lower end, QDA and DT are grouped together with KNN,

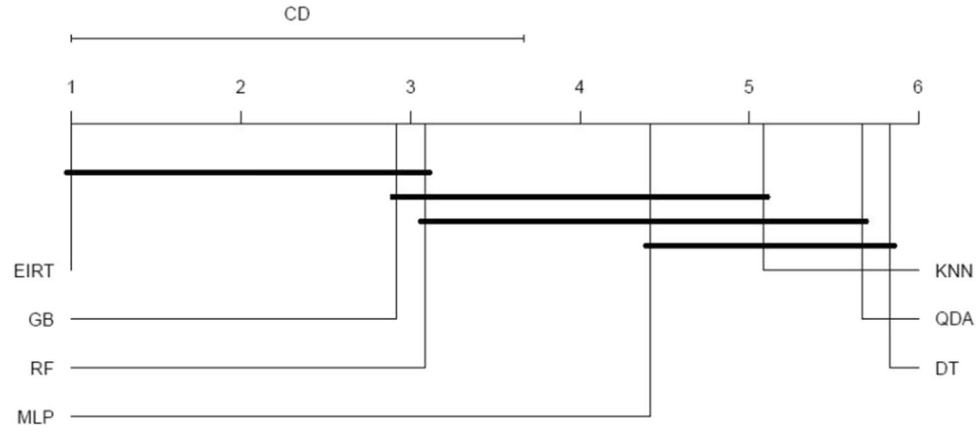


Fig. 8 Results of post hoc tests after Friedman test (AUROC)

with QDA and DT showing the worst performance overall. Note that results of the test based on AUPRs and MSEs shows the same conclusion (see Appendix).

Real-life assessment data examples

Although the purpose of the simulation study was helpful to examine the effect of different aspects of assessments, we acknowledge that the data generation procedure was still constrained by the IRT framework. In this section, we conduct additional experiments based on real-life settings: we employed two educational assessment datasets in university- and national-settings. Same as the simulation study, in each dataset, we conducted a tenfold cross-validation study for three prediction scenarios.

Statistical knowledge assessment data

A first dataset comes from a test from the final grade of the general track of secondary education in Belgium, evaluating a statistical knowledge domain. The dataset consists of the responses of $n = 2044$ students (S) that were assessed on 20 items (I). The students' responses to the items were recorded as dichotomous variables. Specifically, $Y_{ip} = 1$ if the student p has responded to the item i correctly; $Y_{ip} = 0$, otherwise. A set of 22 student-related variables for student properties including status of dyslexia, dyscalculia, AD(H)D, ASS, another language problem, school type, resident area, and so on were incorporated in each method. Similarly, a total of six item-related variables for item properties including question type, attainment target, and so on were incorporated in each method. The categorical variables among those were dummy coded using the *preprocessing* module of the *Scikit-learn* library (Pedregosa et al., 2011) v. 0.23.1 in Python 3.7.7. Additionally, we used Python to impute the few missing values with MICE (van Buuren & Groothuis-Oudshoorn, 2011) through the *IterativeImputer* module in *Scikit-learn*.

National assessment of French data

The national assessment of French data (Denis et al., 2018) consists of the responses of $n = 1950$ students (S) assessed on 22 French listening items (I) administered in primary schools of Flemish region of Belgium. The students' responses to the items were recorded as dichotomous variables. Specifically, $Y_{ip} = 1$ if the student p has responded to the item i correctly and $Y_{ip} = 0$, otherwise. In addition, data from 33 student-related variables were used, including the status of primary language type, dyslexia, dyscalculia, AD(H)D, ASS, another language problem, school type, and so on. Also, 15 item-related variables were used, including attainment type, type of task, visual support, and vocabulary. Similar to the first data example, there are no missing values in the item response data, nor in the item properties data. However, there were missing values in the student; the students with missing values in a majority of the background information were dropped, resulting in a new total of $n = 1918$ students. The remaining missing values in the student properties data were imputed through MICE after a one-hot encoding was applied to the categorical variables.

Results

Table 3 summarizes results of the tenfold CV from the two datasets, including mean values of AUPR, AUROC, and MSE averaged over the ten outer folds and the corresponding SD values (in brackets). For the first dataset, we found that the two tree-ensemble methods i.e., GB and RF perform the best in all three prediction scenarios; EIRM performed well in the next place. On the other hand, in the second dataset,

Table 3 Average AUROC, AUPR, and MSE results from the two datasets

		Real data 1			Real data 2		
		AUROC	AUPR	MSE	AUROC	AUPR	MSE
	EIRM	0.709 (.007)	0.737 (.015)	0.215 (.003)	**0.750 (.009)**	**0.854 (.010)**	**0.182 (.005)**
	RF	0.723 (.009)	0.752 (.016)	**0.210 (.003)**	0.726 (.009)	0.835 (.011)	0.188 (.004)
	GB	**0.725 (.008)**	**0.757 (.014)**	**0.210 (.003)**	0.725 (.008)	0.838 (.010)	0.189 (.005)
New	DT	0.704 (.010)	0.733 (.017)	0.217 (.004)	0.700 (.005)	0.813 (.012)	0.197 (.004)
student	k-NN	0.687 (.009)	0.711 (.013)	0.222 (.003)	0.668 (.011)	0.794 (.012)	0.202 (.005)
scenario	QDA	0.669 (.014)	0.699(.022)	0.256 (.009)	0.666 (.018)	0.797 (.018)	0.351(.038)
	MLP	0.698 (.012)	0.722 (.021)	0.219 (.005)	0.711 (.010)	0.820 (.017)	0.195 (.006)
	EIRM	0.691 (.056)	0.719 (.093)	0.232 (.033)	**0.702 (.054)**	**0.825 (.075)**	0.226 (.061)
	RF	**0.713 (.052)**	**0.745 (.075)**	**0.212 (.020)**	0.653 (.061)	0.784 (.106)	0.225 (.059)
	GB	**0.713 (.047)**	0.741 (.082)	**0.212 (.021)**	0.656 (.055)	0.791 (.096)	0.230 (.060)
New item	DT	0.662 (.045)	0.694 (.088)	0.227 (.021)	0.587 (.050)	0.734 (.112)	0.247 (.070)
scenario	k-NN	0.686 (.048)	0.712 (.079)	0.224 (.017)	0.652 (.027)	0.775 (.101)	**0.216 (.047)**
	QDA	0.608 (.074)	0.646 (.083)	0.298 (.060)	0.545 (.097)	0.722 (.124)	0.432 (.123)
	MLP	0.664 (.051)	0.694 (.105)	0.228 (.031)	0.656 (.050)	0.790 (.095)	0.223 (.058)
	EIRM	0.750 (.004)	0.780 (.005)	0.201 (.002)	**0.777 (.006)**	**0.873 (.005)**	**0.173 (.002)**
Student–item pair	RF	**0.752 (.005)**	**0.784 (.005)**	**0.200 (.002)**	0.730 (.009)	0.841 (.006)	0.187 (.002)
scenario	GB	**0.752 (.004)**	0.783 (.004)	**0.200 (.002)**	0.748 (.008)	0.854 (.007)	0.182 (.002)
	DT	0.712 (.004)	0.746 (.004)	0.214 (.001)	0.701 (.009)	0.815 (.009)	0.196 (.003)
	k-NN	0.717 (.005)	0.744 (.006)	0.212 (.002)	0.697 (.009)	0.817 (.006)	0.197 (.002)
	QDA	0.678 (.009)	0.706 (.006)	0.252 (.005)	0.678 (.011)	0.802 (.007)	0.326 (.020)
	MLP	0.725 (.005)	0.756 (.004)	0.210 (.002)	0.734 (.009)	0.843 (.006)	0.187 (.003)

Best values are indicated in bold and standard deviations in parenthesis

EIRM performed the best in all setting, followed by GB, RF, or MLP in the next places. In other words, the performance ranks were very similar to our simulation results. In both datasets, QDA had the poorest performance.

Among the three prediction scenarios, the best overall performance is seen under the student–item pair scenario as in the simulation result. However, in these real-life assessments, we found that performance on new item scenario is generally poorer than the new student scenario; and the performance was noticeably worse for DT. Considering that there were 6 item-related variables for the first dataset but relatively a greater number of variables in the second dataset (i.e., 15 item-related variables), we found that factors that make the learning task less efficient are both of quantity and quality of the related variables in the training sets.

Conclusion and discussion

The overarching goal of this paper is to examine IRT and ML methods to be able to find ways to obtain more personalized information about student learning. We evaluated the prediction performance in terms of the robustness and accuracy of an EIRM and a range of supervised ML algorithms in the simulated and real-life educational assessment data sets. We found that using student- and item-related background information (explanatory variables) in addition to the student outcome data, we obtain good prediction performance for the cold-start problems, also in situations where no historical data is available for a new student or item. Among the factors that we considered in the simulation, we found that the explanatory power of student- and item-related background information in accounting for variations in student ability and item difficulty has the most impact on the prediction accuracy in any prediction scenario. Therefore, the study recommends that educational researchers and practitioners do not neglect richness of such contextual information about the students and test items when the goal is to predict learning outcomes. What is proposed in this study would be helpful to provide education policy makers and teachers with more accurate group-based statistics capitalizing on rich data in large-scale assessments (e.g., PISA or TIMSS) when the goal is to fine-tune the strategies for building effective teaching and learning environments.

Among the seven prediction methods we used, the simulation study showed that the (unidimensional) EIRM outperformed ML methods in a consistent manner across conditions; the EIRM is more accurate and robust on the whole.

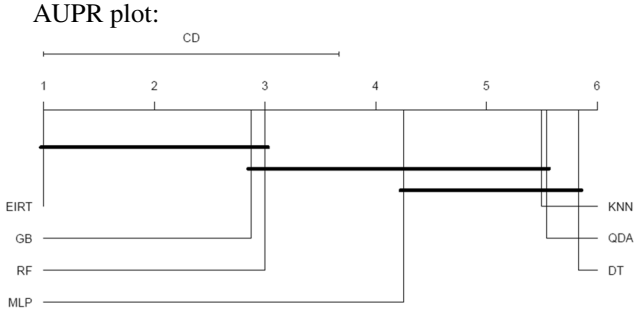
Because the data-generation was carried out by the EIRM (while the student ability parameters was assumed to be multidimensional), however, it must have given the (unidimensional) EIRM a certain level of advantage over the ML models. Even so, it is noteworthy that the strictly "theory-based" EIRM method is also highly competitive, compared to the data-driven ML methods, in solving prediction tasks for the real-life settings of educational assessments. On the other hand, it is also worth noting that some of the ML methods, GB, RF, and MLP, performed as accurately as the EIRM, when the educational assessment data possess properties of IRT. Among the three highly performing ML methods, MLP showed inferior performance to the two tree-ensemble methods. The relatively small number of background information describing the samples (students and items) may explain this phenomenon. Modern deep neural network approaches, albeit effective in general, often fail to meet the related high expectations when it comes to small background information sets (low dimensional feature spaces).

After completing the study, and in view of its limitations, future research with the following methodological challenges would be beneficial. First, regarding the score-point scales, our study focuses on data with dichotomous item responses. We acknowledge that current data we used in our real-life example had a relatively small number of items even compared with other national assessment data. As educational assessments (e.g., TIMSS, PISA) nowadays tend to have more of the constructed-response questions to measure complex problem-solving skills, score-point scales that one assessment data has are likely to be more complex e.g., a combination of dichotomous and polytomous responses. In such cases, it is possible that performance of the ML methods as well as the EIRM could provide an additional perspective on our comparison study. Second, the study can be extended to assessment data from an e-learning environment. In this case, because students have more freedom to access the environment and the students' background may be more diverse, the student cold-start problems should be addressed carefully (e.g., Park et al., 2019). Third, given the excellent performance of EIRM on the one hand and GB or RF on the other hand, it would be interesting to investigate whether their performance can be further boosted by combining them in a hybrid model. Finally, considering more advanced machine learning methods that were specifically designed to learn from interaction data would be interesting. For example, Bi-clustering trees (Pliakos et al., 2018) as well as bi-clustering tree-ensembles (Pliakos & Vens, 2019) are extensions of typical tree-based learning models to the interaction data setting. In such a setting, one has two sets of samples instead of a single one and the output variables to be predicted, often represented as an interaction matrix, define whether two samples interact or not. More specifically, in Pliakos et al. (2018) a bi-clustering tree that integrated features from both sets of samples into a unified learning process was proposed. Next, in Pliakos and Vens (2019) this methodology was extended to tree-ensembles, transferring popular tree-ensemble methods, such as random forests, to the setting of interaction prediction. Model-based collaborative filtering (CF; Bergner et al., 2012) is a ML-based approach that is also capable of estimating parameters for students and items similar to IRT approach. In the domain of CF, a Bayesian probabilistic matrix factorization (Salakhutdinov & Mnih, 2008) seems to be an extensively used technique addressing the student- and item-background information to improve recommendation systems. In addition, (Huang et al., 2020) presents a new deep tabular data modeling architecture for supervised and semi-supervised learning. It is based on self-attention transformers that transform categorical features into robust contextual embedding achieving high prediction performance even in cases with noisy or missing data. In the future, such a model could be utilized to learn from student as well as item related data in order to generate accurate student response predictions.

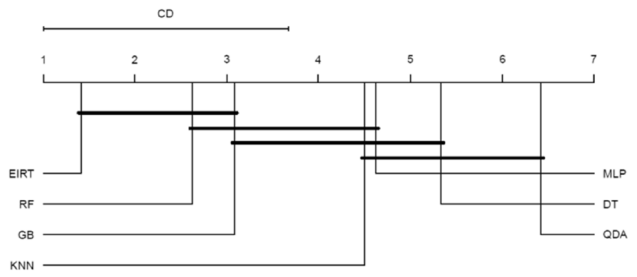
Appendix

Extra figures with post hot tests after the Friedman test being performed on AUPR and MSE are the following:

AUPR plot:



MSE plot:



Acknowledgements This work was carried out within imec's Smart Education research programme, with support from the Flemish government. This research received funding from the Flemish AI Research

Program. Also, this work was supported by the 2021 Research Fund of the University of Seoul for Jinho Kim.

Declaration

Conflicts of interests The authors declared no potential conflicts of interests with respect to the research, authorship and/or publication of this article.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician, 46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Anderson, J. O., Lin, H., Treagust, D. F., Ross, S. P., & Yore, L. D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for international student assessment (PISA). *International Journal of Science and Mathematics Education, 5*(4), 591–614.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version, 1*, 1–17.

Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., & Pritchard, D. E. (2012). Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees (eBook)*. Boca Raton, Florida: Routledge. https://doi.org/10.1201/9781315139470

Calvo, B., & Santafé Rodrigo, G. (2016). Scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal, 8*(1), 248–255.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11*, 2079–2107 http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining*, *13–17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533–559.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer-Verlag.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30.

Denis, J., Carpentier, N., Laenen, I., Willem, L., Janssen, R., & Aesaert, K. (2018). *Peiling Frans in het basisonderwijs – Eindrapport*. Unpublished technical report.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181. https://doi.org/10.1117/1.JRS.11.015020

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 86–92.

Gonzalez, O. (2020). *Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification*. Psychological Methods: Advance online publication. https://doi.org/10.1037/met0000317

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. *Springer Science & Business Media*. https://doi.org/10.1007/978-0-387-84858-7

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science, 349*(6245), 253–255. https://doi.org/10.1126/science.aac4520

Hsia, T. C., Shie, A. J., & Chen, L. C. (2008). Course planning of extension education to meet market demand by using data mining techniques - an example of Chinkuo technology university in Taiwan. *Expert Systems with Applications, 34*(1), 596–602. https://doi.org/10.1016/j.eswa.2006.09.025

Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. arXiv preprint arXiv:2012.06678.

Jiao, H., & Lissitz, R. (2020). What hath the coronavirus brought to assessment? Unprecedented challenges in educational assessment in 2020 and years to come. *Educational Measurement, Issues and Practice, 39*(3), 45–48.

Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement, 80*(4), 726–755.

Kingma, D., & Ba, J. (2017). Adam: A method for stochastic optimization. *ArXiv., 1412*, 6980.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: A decision support system for forecasting students' grades. *Artificial Intelligence Review, 37*(4), 331–344. https://doi.org/10.1007/s10462-011-9234-x

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers and Education, 53*(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Nemenyi, P. (1963). *Distribution-free multiple comparisons*PhD thesis. Princeton University.

Park, J. Y., Joo, S. H., Cornillie, F., et al. (2019). An explanatory item response theory method for alleviating the cold-start problem in adaptive learning environments. *Behav Res, 51*, 895–909. https://doi.org/10.3758/s13428-018-1166-9

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Blondel, M. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2825–2830.

Pliakos, K., Joo, S., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers and Education, 137*, 91–103.

Pliakos, K., Geurts, P., & Vens, C. (2018). Global multi-output decision trees for interaction prediction. *Machine Learning, 107*(8), 1257–1281. https://doi.org/10.1007/s10994-018-5700-x

Pliakos, K., & Vens, C. (2019). Network inference with ensembles of bi-clustering trees. *BMC Bioinformatics, 20*(1), 1–12. https://doi.org/10.1186/s12859-019-3104-y

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1007/bf00116251

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout. *PLoS One, 12*(2). https://doi.org/10.1371/journal.pone.0171207

Salakhutdinov, R., & Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. *In Proceedings of the 25th international conference on Machine learning* (pp. 880–887).

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.

Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: A tutorial. *International journal of applied. Pattern Recognition, 3*(2), 145. https://doi.org/10.1504/ijapr.2016.079050

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R.*Journal of Statistical Software, 45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*(4), 369–386.

Van Der Malsburg, C. (1986). Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the theory of brain mechanisms. In G. Palm & A. Aertsen (Eds.), *Brain theory* (pp. 245–248). Springer-Verlag. https://doi.org/10.1007/978-3-642-70911-1_20

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers Inc.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.

Open Practices Statement Some of the data or materials for the experiments reported here is available at the following GitHub folder (https://github.com/E-IRT-team/E-IRT-ML-comparison). The reader can access our results and scripts that generated simulation datasets. None of the experiments was preregistered.