Article

Journal of Psychoeducational Assessment
2024, Vol. 42(8) 927–954

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/07342829241262861
journals.sagepub.com/home/jpa
Sage

Combining Statistical and Machine Learning Methods to Identify Predictors of Brazilian Students' Proficiency in PISA 2018

Pedro San Martin Soares[1,2]

Abstract

Brazil's education system lags behind international standards, with two-fifths of students scoring below the minimum level of proficiency in mathematics, science, and reading. Thus, this study combined machine learning with traditional statistics to identify the most important predictors and to interpret their effects on proficiency in the PISA 2018 mathematics, science, and reading tests. Predictors encompassed a wide range of variables, sociodemographic characteristics, teaching and learning processes, and non-cognitive skills. The outcome of the present study was proficiency in mathematics, science, and reading. PISA proficiency levels were grouped into "low proficiency" and "proficient" categories, using a classification system commonly employed in PISA reports. Using random forest analysis, a machine learning method, I compared the importance of predictors for proficiency in mathematics, science, and reading. I then adjusted multilevel logistic regression analyses to investigate the relationship between the top predictors and the outcomes. Among the top predictors for the three outcomes identified, annual household income, parents' highest occupational status, and early childhood education and care were positively associated with proficiency in mathematics, science, and reading, while grade repetition and additional instruction were negatively associated with these outcomes. These findings urge Brazilian policymakers and educators to prioritize initiatives that strengthen early childhood programs, minimize grade repetition, and promote effective learning strategies.

Keywords

programme for international student assessment, international student assessment, adolescence, student achievement

[1]Universidade Federal de Pelotas, Pelotas, Brazil
[2]State Health Department of Rio Grande do Sul, Porto Alegre, Brazil

Corresponding Author:
Pedro San Martin Soares, Universidade Federal de Pelotas, Marechal Deodoro, 1160 e 3° andar, CEP: 96020 e 220, Pelotas 96010-610, Brazil.
Email: pedrosmsoares@hotmail.com

Introduction

International education assessments are large-scale, standardized tests designed to measure student performance across different countries and cultures. Examples of these assessments include Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS). These assessments provide valuable data for a variety of purposes, including comparative international benchmarking, identifying global trends in education, fostering cross-national collaboration, informing policy development, and driving improvements and accountability in education systems. They provide countries with a way to compare their educational outcomes, identify strengths and weaknesses, and track progress over time. Insights into global trends and factors associated with high achievement can inform the development of effective intervention strategies.

Identifying predictors of educational performance in large-scale assessments provides an important tool for the effective planning of educational systems. By understanding these factors, educators and policymakers can identify strengths and weaknesses, enabling targeted improvements in policies and practices. This data-driven approach empowers managers to make informed decisions, ensuring that resources and interventions are effectively allocated. Identifying predictors facilitates personalized programs to meet specific student needs, such as focusing initiatives on disadvantaged groups if socioeconomic background plays a significant role. Furthermore, they offer valuable information for broader education reforms, potentially encompassing curriculum changes, teacher development, or resource allocation. This knowledge enables educators to address issues of equity, highlighting how factors such as socioeconomic status, gender, or region influence performance. This allows for policies that ensure every student has the best possible opportunity for success.

PISA Proficiency Levels

Among the many international assessments, the PISA has been one of the most widely used to provide valuable information about the strengths and weaknesses of educational systems around the world (Takayama, 2008). The PISA provides a broader picture of how well education systems prepare students for real-world challenges, while other assessments, such as TIMSS and PIRLS, offer more focused evaluations of specific curriculum content within particular subject areas (Kell & Kell, 2014). Originally developed by the Organization for Economic Cooperation and Development (OECD), PISA assessments of mathematics, reading, and science are administered internationally every 3 years with 15-year-old students of both OECD member and participating non-member countries, with participation increasing from 35 countries in 2000 to 77 countries in 2018 (OECD., 2019).

PISA assessment scores are scaled to achieve an OECD average score of 500, with a standard deviation of 100 (OECD., 2019). This means that about two-thirds of students in OECD countries are expected to score between 400 and 600 points. To disseminate results and provide pedagogical interpretation, PISA groups students' test scores into six proficiency levels. Each level reflects a specific set of skills and knowledge, with Level 2 considered the baseline proficiency level for full participation in society. Students achieving Level 2 or above, with minimum scores of 420 in mathematics, 410 in science, and 407 in reading, are considered proficient. While the average PISA score is helpful for international comparisons and rankings, proficiency levels are more relevant for planning and implementing educational policies, as they allow countries to pinpoint areas of strength and weakness in their education systems.

Predictors of PISA Performance

Reflections on the importance of PISA data for investigating educational performance show that its micro-level data provide information on individual variables known to be associated with educational performance. These factors can be broadly categorized as student sociodemographic characteristics, teaching and learning processes, and non-cognitive skills. Sociodemographic characteristics like socioeconomic status and early childhood education have been consistently positively associated with student performance, while age and grade retention tend to have a negative association with it (Gamazo et al., 2018; Karakolidis et al., 2016; Pholphirul, 2017). Regarding sex, PISA results across participating countries indicate that girls tend to outperform boys in reading, while boys tend to outperform girls in mathematics (OECD, 2019b). Teaching and learning processes, including school climate, learning time, teacher support, and frequency of computer use, also influence PISA scores (Gómez-Fernández & Mediavilla, 2021; Kuger et al., 2016; Lazarevic & Orlic, 2018; She et al., 2019; Touron et al., 2018). Non-cognitive skills like motivation, self-efficacy, and strategic learning approaches have been shown to positively associate student performance, particularly in reading (Aksu & Güzeller, 2016; Cheung et al., 2014; She et al., 2019; Touron et al., 2018).

Combining Statistical and Machine Learning Approaches to Data Analysis in Educational Research

Studies analyzing PISA data employ a wide range of statistical methods. While a comprehensive description of these methods is beyond the scope of this paper, traditional statistics such as multilevel regression analysis, structural equation modeling, and ANCOVA are the most common (Gamazo & Martínez-Abad, 2020). Although such methods are highly simplistic and interpretable, they have several limitations that may lead to inaccurate predictions and conclusions (Hindman, 2015). Traditional statistical methods can struggle to effectively analyze data comprising a large number of participants and predictors. Empirical studies typically select specific factors based on theory or evidence. Even with more sophisticated methods applied to extensive datasets, such as multilevel linear models for PISA data, researchers often still narrow their focus to a subset of predictors. Including too many predictors simultaneously in the model risks overfitting, which occurs when the established model is unable to generalize beyond the data used to build it. Additionally, the assumption of linearity between variables in most of these methods, combined with the numerous and varied predictors that influence academic performance (e.g., students' demographic, personality, psychological, and environmental variables), can lead to a failure to recognize potential nonlinearities (Guo et al., 2015; Strobl et al., 2009).

In recent years, researchers have begun employing machine learning approaches in studies using PISA data (Gamazo & Martínez-Abad, 2020; Immekus et al., 2022). Machine learning is an umbrella term for a wide variety of models and strategies focused on algorithmic modeling that often involves data preprocessing, such as splitting data into training and testing sets. It also involves variable selection, which may be less theory-driven than in traditional statistics, and model evaluation, which prioritizes predictive power over explanatory clarity (Royal Society, 2017). Machine learning models are generally grouped into unsupervised or supervised models. Unsupervised models discover unknown patterns and groupings within data without reference to any specific outcome (e.g., identifying that oranges and beach balls share a round shape). Supervised models, in contrast, typically predict outcomes, similar to traditional statistical methods like logistic regression. Supervised learning can be further categorized as either classification models (predicting discrete categories) or regression models (predicting continuous values; Nasteski, 2017).

For over two decades, researchers have explored the potential of combining traditional statistics and machine learning (Friedman, 1997). Despite both offering valuable tools for analyzing PISA data, studies combining these methodologies remain scarce. Traditional statistics, while effective for computing the magnitude of relationships among variables through hypothesis testing, were initially developed for smaller datasets with fewer variables (Bzdok et al., 2018). As the number of variables and sample size increase, traditional statistical approaches may lead to an overrepresentation of significant results (Martínez-Abad et al., 2020). Machine learning approaches may complement traditional statistics and help address some of their limitations. Recently, some research teams have adopted a methodology that first uses supervised machine learning classification methods to determine the importance of a wide range of predictors (Lezhnina & Kismihók, 2022; Wang et al., 2023). Then, traditional statistics are applied to this refined subset of predictors (identified as important in the machine learning stage) to estimate predictive associations based on known statistical distributions. This two-step process leverages the strengths of both approaches: traditional statistics provide explanatory power and clear interpretability, while machine learning offers accurate prediction.

Random Forest Algorithm (RFA)

The random forest algorithm (RFA), one of the key supervised machine learning classification algorithms, is increasingly used in combination with traditional statistics because it offers greater robustness and accuracy for classification tasks (Williamson et al., 2021), often outperforming logistic regression (Breiman, 2001b; Couronné et al., 2018). This approach relies on decision trees to divide datasets based on variables that maximize information gain at each split. Specifically, the RFA randomly builds regression trees based on random subsets of the observation and creates each split of a tree based on a random subset of candidate variables (Breiman, 2001a). The process is combined with bootstrapping procedures to iteratively minimize the predictive error (Yarkoni & Westfall, 2017). A key advantage of random forest over traditional statistics is its ability to model both individual predictor effects and the complex multivariate interactions among predictors (Breiman, 2001a). Additionally, it can be applied in situations where predictors are highly correlated. Such correlation might violate collinearity assumptions in traditional statistical methods, but the random forest handles this by randomly selecting variables within each tree.

Brazil's Performance in the PISA

Brazil's education quality lags behind international standards, as indicated by the PISA released in 2018. The data revealed that Brazil ranked 57th out of 77 participating countries in reading, 70th in mathematics, and 66th in science (OECD., 2019). While 50% of Brazilian students achieved reading proficiency (Level 2 or above on the proficiency scale), meaning they can, at a minimum, identify main ideas and interpret explicitly stated information in texts, this falls short of the OECD average of 77%. The gap widens further in mathematics, where only 32% of Brazilian students were proficient, compared to the OECD average of 76%. Students achieving proficiency in this subject possess the fundamental skills to apply basic mathematical concepts to real-world situations, such as comparing distances or converting currencies. This disparity is even more striking when compared to better-performing regions such as Beijing, Shanghai, Jiangsu, and Zhejiang in China, where 98% of students achieved proficiency. Science proficiency also reveals a significant gap, with only 45% of Brazilian students reaching Level 2 or above, compared to the OECD average of 78%. Students reaching this benchmark possess a foundational understanding of scientific phenomena and can use this knowledge to evaluate simple conclusions based on given data (OECD., 2019).

In 2018, around 43% of Brazilian students did not achieve proficiency in all three subjects (below Level 2 on the proficiency scale) (OECD., 2019). Students performing below proficiency on the PISA scale face considerable challenges in navigating academic and real-world demands. In reading, they struggle to comprehend texts that go beyond basic vocabulary and sentence structures, limiting their access to information and hindering their ability to think critically. Mathematics poses hurdles when multiple steps or abstract concepts are involved, making it difficult for them to solve complex problems or interpret data effectively. In science, they grapple with understanding intricate concepts, applying them to new situations, and critically evaluating scientific evidence—all essential for informed decision-making. These difficulties highlight the need to identify the various factors influencing the achievement of Level 2 or above and to formulate effective policies and strategies that improve the educational system, thus boosting the country's social and economic advancement.

The Present Study

The PISA is considered among the most advanced international assessments to date, encompassing roughly nine-tenths of the world economy. Improving Brazilian students' proficiency in the subjects assessed by the PISA has the potential to reduce the performance gap with other countries and ensure Brazil's competitiveness on the global stage. Formulating public policies and interventions to enhance school performance indices presents a multifaceted challenge, as school performance is influenced by a variety of factors, including individual, family, and school-based determinants. One of the primary challenges in improving educational outcomes, especially in low- and middle-income countries with limited resources for public policy investments, is identifying the key variables that influence educational performance and should be prioritized for intervention. By combining traditional statistical methods with a machine learning approach to compare such a wide range of predictors, we can enhance our understanding of the factors influencing educational performance in Brazil. Thus, this study combined machine learning with traditional statistics to identify the most important predictors among 51 variables—including sociodemographic characteristics, teaching and learning processes, and non-cognitive skills—and to interpret their effects on proficiency in the PISA 2018 mathematics, science, and reading tests.

Methods

Participants

Public use data files from the PISA administered in 2018 were analyzed. The PISA is a cross-sectional complex survey involving multistage sampling, unequal sampling probabilities, and stratification. The population of this study were 15-year-old students and schools from Brazil participating in PISA 2018. The sample was composed of 10,691 15-year-old students and 597 schools. Detailed descriptions of the PISA methodology have been published elsewhere (OECD, 2019a).

Measurements

PISA Proficiency Levels. The PISA uses standardized tests to assess the reading, mathematics, and science skills of 15-year-old students in participating countries. The tests are designed to measure how well students can apply their knowledge to real-life problems, rather than just memorize information. The tests consist of a combination of multiple-choice and open-ended questions and are administered in both digital and paper-based formats (OECD, 2019a).

The scores for the reading, mathematics, and science tests in the PISA are calculated based on the number of questions that students answered correctly, with more difficult questions receiving a higher score (OECD, 2019a). The scores are computed according to Item Response Theory (IRT) and standardized with a mean of around 500 and standard deviation of around 100 (OECD, 2019c). To interpret these standardized scores and facilitate international comparison, the PISA employs a proficiency scale. This scale, divided into six levels ranging from 1 (lowest) to 6 (highest), outlines the specific skills and knowledge students should demonstrate at each level. This standardized classification system enables meaningful comparisons of student performance across different countries.

In the present study, PISA proficiency levels were grouped into two categories: "Low Proficiency" (below Level 2) and "Proficient" (Level 2 or above). This decision to dichotomize PISA proficiency levels aligns with practices established in PISA reports and research (OECD, 2019b) and is supported by evidence linking the attainment of Level 2 to reduced school dropout rates and increased access to higher education (Knighton, 2006; OECD, 2007; Schleicher, 2007).

Predictors

The individual predictors were obtained from PISA 2018 context questionnaires answered by students. These questionnaires contain a wide range of sociodemographic, economic, and educational information related to student outcomes (OECD, 2019a). Several measures reflect indices summarizing responses to a series of related questions. The present study included 51 predictors (representing 5 of 7 modules administered to students). These predictors covered sociodemographic characteristics, teaching and learning processes, and non-cognitive skills. Detailed descriptions of how each scale index was constructed are available in the PISA 2018 Technical Report.

Sociodemographic characteristics

The analyses included 11 sociodemographic factors: sex (male/female), age (in complete years), annual household income, household possessions, parents' highest occupational status, the highest maternal and paternal education, early childhood education and care (in complete years; less than one year, between one and two, more than two years), grade relative to the country's modal grade (above, at, or below modal grade), grade repetition (no/yes), and number of educational transitions (none, one, two, three, or four times).

Annual household income was collected in local currency and converted into multiples of the Brazilian monthly minimum wage at the time of data collection, as it is a commonly used parameter in Brazil to measure income (Mostafa & Silva, 2007). Household possessions were assessed using a scale that evaluated ownership of 16 household items, including three country-specific items that were considered appropriate measures of household wealth in the country's context. Additionally, the number of belongings and books at home was included in the scale. The total scale score was categorized into quartiles for the analysis.

Parents' highest occupational status was determined based on the International Socio-Economic Index of Occupational Status (ISEI; OECD, 2019a). The ISEI score used was the higher of either parent's ISEI score or the only available parent's ISEI score. ISEI scores range from 0 to 100, with higher scores indicating higher occupational status. Parental education levels were based on the International Standard Classification of Education (Unesco, 1997), which is a seven-point scale that ranks education levels in ascending order of the highest level completed. Indices on parental education were constructed by recoding educational qualifications into the following categories: None; ISCED level 1 (primary education); ISCED level 2 (lower secondary); ISCED level 3B or 3C (vocational/pre-vocational upper secondary); ISCED level 3A (general upper secondary) and/or ISCED level 4 (non-tertiary post-secondary); ISCED level 5B (vocational tertiary); and ISCED level 5A and/or ISCED level 6 (theoretically oriented tertiary and post-graduate).

Teaching and Learning Processes

Fifteen factors related to teaching and learning processes were included in the analysis. These factors encompass frequency of entire days of school classes skipped (without authorization in the 2 weeks prior; never, one or two times, three or four times, and five or more times), frequency of skipped school classes (without authorization in the 2 weeks prior; never, one or two times, three or four times, and five or more times), frequency of lateness for school (in the 2 weeks prior; never, one or two times, three or four times, and five or more times), additional instruction (such as enrichment or remedial lessons for language, mathematics, science, or foreign language tests; no/yes), learning time at school (in mathematics, language tests, and learning time; minutes per week). Additionally, 10 scales related to Information and Communication Technology (ICT) were included: ICT use outside of school for leisure scale, ICT use outside of school for learning scale, ICT use at school in general, interest in ICT, perceived ICT competence, perceived autonomy related to ICT use, ICT as a topic in social interaction, information about careers, information about the labor market provided by the school, and information about the labor market provided outside of school. Each ICT scale used a self-reported, Likert-type format on a logit scale, with 0 representing the OECD average.

Non-Cognitive Skills

The analysis included scores on 25 non-cognitive skill scales. The motivational scales assessed the following factors: student's expected occupational status, disciplinary climate in test language lessons, teacher support in test language lessons, teacher-directed instruction, perceived feedback, parents' emotional support as perceived by the student, teacher's stimulation of reading engagement as perceived by the student, adaptation of instruction, perceived teacher's interest, joy/liking for reading, perception of competence in reading, perception of difficulty in reading, perception of competitiveness at school, perception of cooperation at school, attitude toward school learning activities, competitiveness (dispositional desire to outperform others), work mastery (dispositional desire to work hard to master tasks), general fear of failure, eudaemonia (sense of meaning and purpose in life), positive affect, resilience, mastery goal orientation, sense of belonging to school, and student's experience of being bullied. The metacognitive scales assessed understanding and remembering, summarizing, and assessing credibility.

Analysis

Analyses were performed separately for each of the outcomes using Python software version 3.10.12 (Python Software Foundation), and statistical significance was set at 0.05. For descriptive analysis, we used chi-square to compare proportions and ANOVA to compare means between the groups.

To obtain a more detailed picture of relationships between predictors and outcomes, a two-step analysis was followed. First, a random forest algorithm was used to identify the most important predictors. Then, considering the nested structure of the data, multilevel logistic regression was employed to calculate effect sizes and interpret the results. Previous studies have indicated that single-level random forests produce results similar to multilevel (or mixed-effect) random forests and can effectively handle hierarchical data (Mangino & Finch, 2021).

Therefore, this study employed a single-level (or fixed effect) random forest algorithm. However, research demonstrates that increasing group size benefits random forests more significantly. In contrast, estimation models like multilevel logistic regression benefit primarily from a larger number of groups (Afshartous & de Leeuw, 2005; Breiman, 2001a). Considering the substantial group sizes (the number of students per school, ranging from 1 to 30) and the large number of schools in this study, multilevel logistic regression was the choice for the next step. This approach aligns with the combination of statistical and machine learning methods used in previous studies analyzing PISA data (Lezhnina & Kismihók, 2022; Peng et al., 2023; Wang et al., 2023).

The data were split into analysis groups (A1 = random forest analysis and A2 = multilevel logistic regression) using Python's built-in random sampling function, resulting in two subsamples: A1 ($n = 5345$) and A2 ($n = 5346$). To handle missingness, multiple imputation by chained equations was used to impute complete datasets for each of these two subsamples, using all variables included in the analyses. All analyses were performed separately for each of the academic outcomes: science, math, and reading.

RFA

To estimate the importance of variables and variables' ranks, an RFA was used. This method combines multiple decision trees to improve the accuracy and robustness of predictions. Two key hyperparameters (parameters defined before the learning process begins) in the RFA are the number of trees (ntree), which determines how many decision trees are constructed, and the number of pre-sampled variables per node (mtry), which controls how many variables are randomly selected and considered at each split within a tree. Increasing ntree generally improves performance by reducing overfitting, but excessive trees lead to diminishing returns and higher computational costs. For mtry, lower values often lead to more correlated, potentially overfit trees, while higher values result in less correlated, potentially underfit trees. Ideal ntree and mtry values depend on the specific problem and dataset. A common starting point for classification tasks is 100–500 trees and the square root of the total number of variables for mtry. However, hyperparameter optimization techniques can be used to identify the best combination of these values for optimal model performance and training time.

GridSearchCV is a hyperparameter optimization technique that systematically tests all possible combinations of values provided for each hyperparameter. It utilizes cross-validation to evaluate the performance of each combination. K-fold cross-validation divides the dataset into K folds (subsets), trains the model on K-1 folds, and evaluates it on the remaining fold. This process is repeated K times, with each fold serving as the test set once. In each round of cross-validation, GridSearchCV tests different combinations of ntree and mtry, averaging the performance for each combination. Finally, it selects the combination of hyperparameters that yields the best average performance, aiding in the identification of the most accurate and robust random forest model for your data. A common choice for K is 5 or 10, as these values provide a good balance between bias and variance when estimating model performance.

The sklearn.ensemble library for Python was utilized to implement the RFA, enabling the option for mixed-type predictors (Pedregosa et al., 2011). Specifically, the RFA was applied to subsample A1 using the RandomForestClassifier function. The following steps were used for each outcome variable independently: first, the dataset was randomly divided into an 80% training set and a 20% test set, ensuring a similar distribution of the outcome variable across both. Second, to determine the optimal ntree and mtry hyperparameters, the GridSearchCV function with 5-fold cross-validation was used. Various hyperparameter values were explored (ntree = 100−1200 and mtry = 5−10). Third, RandomForestClassifier hyperparameters were fine-tuned. Finally, these steps were repeated 100 times with different random seeds. Using different seeds diversifies individual trees and reduces the model's overall variance, ultimately improving its performance.

The mean decrease in accuracy (MDA) metric was used to assess the importance of each predictor (Strobl et al., 2007, 2009). This metric quantifies how much the model's predictive accuracy decreases when a predictor's values are randomly permuted. First, a baseline accuracy is established using the model with all predictors included. Then, the values of a specific predictor are shuffled, effectively randomizing it and removing its relationship to the outcome. The model's accuracy is recalculated with the shuffled predictor. The MDA for that predictor is the difference between the baseline accuracy and the accuracy with the permuted predictor. MDA values range from 0.000 to 1.000, where a high MDA indicates a greater decrease in accuracy when the predictor is permuted, suggesting that the variable is a strong predictor of the outcome. In the present study, variables were classified according to the average MDA of 100 repetitions, and the 10 top-ranked predictors are shown in the variable importance analysis.

Out-of-bag (OOB) accuracy estimation was used to assess the model's generalization performance. OOB observations are those not sampled during the construction of each individual decision tree. Each tree generates predictions for its own OOB observations, samples on which it was not trained. These predictions are then compared with the actual values of the OOB observations. Overall OOB accuracy is calculated as the average proportion of correct predictions across all trees in the forest for their respective OOB samples (scaled from 0% to 100%, where 100% indicates perfect accuracy). This approach provides an unbiased estimate of model accuracy without the need for a separate validation set, as the OOB observations act as an integrated test set for each tree (Janitza & Hornung, 2018).

The RFA demonstrates flexibility when dealing with continuous, categorical, and ordinal variables in building its decision trees. For continuous variables (such as learning time), the algorithm selects optimal cutoff points at each node, dividing the data into two groups based on these values. For categorical and ordinal variables (such as household possessions and mother's education), the process involves evaluating different divisions based on category levels, selecting the one that best separates the classes, which may result in more than two subsets. Internally, the algorithm automatically handles encoding and may use techniques such as one-hot encoding (creating binary dummy variables for categorical variables) or ordinal encoding (assigning numerical values to categories based on their inherent order) to represent these types of variables (Wright & König, 2019).

An RFA stratified by sex was conducted; however, the top 10 predictors remained identical for the three outcomes across sexes. Consequently, the sex variable was included as a predictor, and the results are presented combined. The Python code and all instructions necessary to repeat the procedures described are included in the Supplementary Material.

Multilevel Logistic Regression

After selecting the most important predictors through the RFA, adjusted multilevel logistic regression analyses were conducted to calculate the odds ratios of the relationship between the predictors and the outcomes, while accounting for the potential influence of multiple students in the same school. Since students from the same school may share similar environmental and social factors that could influence their outcomes, we treated student measures as the first level, which was modeled as fixed effects, and school as the second level, which was modeled as random effects, within which students are nested. Additionally, the models were graphically assessed for normality of residuals and homoscedasticity, and the correlation between continuous predictors was evaluated (Supplementary Material). To compare statistical differences between ordinal categories (such as mother's education), heterogeneity and linear trend tests were used, reporting the category with the lower p-value. All predictors selected in the previous step through the RFA were simultaneously included in the adjusted multilevel logistic regression models.

The improvement of utilizing schools as a random factor in the model compared to a usual logistic regression model was evaluated based on changes in the −2loglikelihood. Under the null hypothesis, this metric follows a $\chi 2$ distribution with k degrees of freedom, where k represents the number of supplementary parameters in the more complex model. Additionally, the intraclass correlation coefficient (ICC) was estimated to quantify the percentage of variability in the outcomes that can be explained by differences between schools.

The metric used to evaluate the set of predictors included in the adjusted multilevel logistic regression models was the area under the receiver operating characteristic curve (AUC-ROC). The receiver operating characteristic curve is a probability curve, plotted with the true positive rate (sensitivity) on the y-axis against the false positive rate on the x-axis. The AUC-ROC measures the model's ability to distinguish between individuals with and without the outcome (e.g., proficient or not proficient) based on their predicted probabilities. A perfect classifier has an AUC-ROC of 100%, while a random classifier, where the true positive rate equals the false positive rate, has an AUC-ROC of 50% (Wagner & Merlo, 2014). AUC-ROC values were based on fixed effects and posterior means of random effects.

Results

Sample Characteristics

Distribution of sample study characteristics by high performance of mathematics, science, and reading is provided in Table 1.

RFA

The RFA tuning parameters are provided in the Supplemental Materials. When run with the best parameter values found for each outcome, the RFAs yielded an out-of-bag accuracy of 87.2% (95% CI: 87.2%–87.3%; ntree = 10; mtry = 800) for mathematics, 86.7% (95% CI: 86.6%–86.8%; ntree = 10; mtry = 200) for science, and 86.9% (95% CI: 86.8%–87.0%; ntree = 7; mtry = 800) for reading.

The RFA results for the 10 predictors with the highest variable importance are presented in Table 2, with the full list available in the Supplemental Materials. For mathematics proficiency, the most important predictors were annual household income, followed by metacognition summarizing subscale, metacognition assessing credibility subscale, parents' highest occupational status, metacognition understanding and remembering subscale, additional instruction, early childhood education and care, frequency of skipped school classes, grade repetition, and learning time. The top predictors were highly similar for science and reading proficiency, in descending importance order: metacognition assessing credibility subscale, additional instruction, metacognition summarizing subscale, annual household income, metacognition understanding and remembering subscale, parents' highest occupational status, perceived autonomy related to ICT use, household possessions, grade repetition, and early childhood education and care.

Multilevel Logistic Regression Analysis

The results of the multilevel logistic regression analysis are presented in Table 3. Annual household income and parents' highest occupational status were positively associated with

Table 1. Distribution of Sample Characteristics by Proficiency in Mathematics, Science, and Reading, PISA 2018 (N = 10.691).

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
Sociodemographic characteristics										
Sex				<0.001[a]			<0.001[a]			<0.001[a]
Female	5478 (51.2)	4722 (86.2)	756 (13.8)		4633 (84.6)	845 (15.4)		4432 (80.9)	1046 (19.1)	
Male	5213 (48.8)	4237 (81.3)	976 (18.7)		4253 (81.6)	960 (18.4)		4407 (84.5)	806 (15.5)	
Age (mean, SD)	15.9 (0.3)	15.9 (0.3)	15.9 (0.3)	<0.001[b]	15.9 (0.3)	15.9 (0.3)	<0.001[b]	15.9 (0.3)	15.9 (0.3)	<0.001[b]
Annual household income (in minimum wages; missing = 3.247)				<0.001[a]			<0.001[a]			<0.001[a]
Up to 10	4537 (60.9)	3770 (83.1)	767 (16.9)		3766 (83.0)	771 (17.0)		3725 (82.1)	812 (17.9)	
More than 10 to 20	2174 (29.2)	1319 (60.7)	855 (39.3)		1320 (60.7)	854 (39.3)		1350 (62.1)	824 (37.9)	
More than 20	733 (9.9)	412 (56.2)	321 (43.8)		413 (56.3)	320 (43.7)		409 (55.8)	324 (44.2)	
Household possessions (quartile; missing = 161)				<0.001[a]			<0.001[a]			<0.001[a]
First	2633 (25.0)	2520 (95.7)	113 (4.3)		2494 (94.7)	139 (5.3)		2480 (94.2)	153 (5.8)	
Second	2632 (25.0)	2422 (92.0)	210 (8.0)		2396 (91.0)	236 (9.0)		2359 (89.6)	273 (10.4)	
Third	2633 (25.0)	2207 (83.8)	426 (16.2)		2184 (82.9)	449 (17.1)		2157 (81.9)	476 (18.1)	
Fourth	2632 (25.0)	1650 (62.7)	982 (37.3)		1651 (62.7)	981 (37.3)		1687 (64.1)	945 (35.9)	
Parents' highest occupational status (ISEI; missing = 451)	41.9 (22.8)	37.5 (21.1)	53.54 (23.0)	<0.001[b]	37.5 (21.2)	53.5 (22.9)	<0.001[b]	37.9 (21.4)	52.7 (23.0)	<0.001[b]

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
Mother's education (ISCED; missing = 451)				<0.001[a]			<0.001[a]			<0.001[a]
None	1087 (10.6)	1042 (95.9)	45 (4.1)		1038 (95.5)	49 (4.5)		1020 (93.8)	67 (6.2)	
ISCED 1	1154 (11.3)	1059 (91.8)	95 (8.2)		1055 (91.4)	99 (8.6)		1056 (91.5)	98 (8.5)	
ISCED 2	1559 (15.2)	1396 (89.5)	163 (10.5)		1365 (87.6)	194 (12.4)		1349 (86.5)	210 (13.5)	
ISCED 3B, C	2508 (24.5)	2085 (83.1)	423 (16.9)		2044 (81.5)	464 (18.5)		2005 (79.9)	503 (20.1)	
ISCED 3A, 4	349 (3.4)	308 (88.3)	41 (11.7)		306 (87.7)	43 (12.3)		311 (89.1)	38 (10.9)	
ISCED 5B	2576 (25.2)	1786 (69.3)	790 (30.7)		1781 (69.1)	795 (30.9)		1827 (70.9)	749 (29.1)	
ISCED 5A, 6	1007 (9.8)	847 (84.1)	160 (15.9)		862 (85.6)	145 (14.4)		843 (83.7)	164 (16.3)	
Father's education (ISCED; missing = 806)				<0.001[a]			<0.001[a]			<0.001[a]
None	1369 (13.8)	1294 (94.5)	75 (5.5)		1278 (93.4)	91 (6.6)		1271 (92.8)	98 (7.2)	
ISCED 1	1350 (13.7)	1230 (91.1)	120 (8.9)		1208 (89.5)	142 (10.5)		1209 (89.6)	141 (10.4)	
ISCED 2	1416 (14.3)	1248 (88.1)	168 (11.9)		1243 (87.8)	173 (12.2)		1194 (84.3)	222 (15.7)	
ISCED 3B, C	2123 (21.5)	1688 (79.5)	435 (20.5)		1663 (78.3)	460 (21.7)		1653 (77.9)	470 (22.1)	
ISCED 3A, 4	338 (3.4)	282 (83.4)	56 (16.6)		279 (82.5)	59 (17.5)		278 (82.2)	60 (17.8)	

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
ISCED 5B	2016 (20.4)	1384 (68.7)	632 (31.3)		1383 (68.6)	633 (31.4)		1424 (70.6)	592 (29.4)	
ISCED 5A, 6	1273 (12.9)	1077 (84.6)	196 (15.4)		1075 (84.4)	198 (15.6)		1055 (82.9)	218 (17.1)	
Early childhood education and care (in years; mean, SD; missing = 3082)				<0.001[a]			<0.001[a]			<0.001[a]
Less than 1 year	383 (5.0)	346 (90.3)	37 (9.7)		344 (89.8)	39 (10.2)		342 (89.3)	41 (10.7)	
Between 1 and 2 years	2674 (35.2)	1853 (69.3)	821 (30.7)		1847 (69.1)	827 (30.9)		1883 (70.4)	791 (29.6)	
More than 2 years	4552 (59.8)	3395 (74.6)	1156 (25.4)		3415 (75.0)	1136 (25.0)		3400 (74.7)	1152 (25.3)	
Grade compared to modal grade in country				<0.001[a]			<0.001[a]			<0.001[a]
Above modal grade	5864 (54.9)	5310 (90.6)	554 (9.4)		5304 (90.5)	560 (9.5)		5280 (90.0)	584 (10.0)	
At modal grade	4608 (43.1)	3490 (75.7)	1118 (24.3)		3437 (74.6)	1171 (25.4)		3408 (74.0)	1200 (26.0)	
Below modal grade	219 (2.0)	159 (72.6)	60 (27.4)		145 (66.2)	74 (33.8)		151 (68.9)	68 (31.1)	
Grade repetition (missing = 253)				<0.001[a]			<0.001[a]			<0.001[a]
No	7211 (69.1)	5567 (77.2)	1644 (22.8)		5501 (76.3)	1710 (23.7)		5460 (75.7)	1751 (24.3)	
Yes	3227 (30.9)	3140 (97.3)	87 (2.7)		3132 (97.1)	95 (2.9)		3131 (97.0)	96 (3.0)	

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
Number of changes in educational biography (missing = 2267)				<0.001[a]			<0.001[a]			<0.001[a]
None	2791 (33.1)	2139 (76.6)	652 (23.4)		2095 (75.1)	696 (24.9)		2099 (75.2)	692 (24.8)	
One	2440 (29.0)	1939 (79.5)	501 (20.5)		1942 (79.6)	498 (20.4)		1890 (77.5)	550 (22.5)	
Two	1917 (22.8)	1656 (86.4)	261 (13.6)		1656 (86.4)	261 (13.6)		1664 (86.8)	253 (13.2)	
Three	704 (8.4)	591 (83.9)	113 (16.1)		575 (81.7)	129 (18.3)		588 (83.5)	116 (16.5)	
Four	572 (6.8)	507 (88.6)	65 (11.4)		499 (87.2)	73 (12.8)		496 (86.7)	76 (13.3)	
Teaching and learning processes										
Frequency of entire days of school classes skipped* (missing = 3648)				<0.001[a]			<0.001[a]			<0.001[a]
Never	3569 (50.7)	2750 (77.1)	819 (22.9)		2731 (76.5)	838 (23.5)		2726 (76.4)	843 (23.6)	
One or two times	2603 (37.0)	2160 (83.0)	443 (17.0)		2124 (81.6)	479 (18.4)		2104 (80.8)	499 (19.2)	
Three or four times	508 (7.2)	450 (88.6)	58 (11.4)		432 (85.0)	76 (15.0)		434 (85.4)	74 (14.6)	
Five or more times	363 (5.2)	335 (92.3)	28 (7.7)		333 (91.7)	30 (8.3)		331 (91.2)	32 (8.8)	

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
Frequency of skipped school classes* (missing = 3699)				<0.001[a]			<0.001[a]			<0.001[a]
Never	3488 (49.9)	2609 (74.8)	879 (25.2)		2553 (73.2)	935 (26.8)		2574 (73.8)	914 (26.2)	
One or two times	2599 (37.2)	2217 (85.3)	382 (14.7)		2197 (84.5)	402 (15.5)		2154 (82.9)	445 (17.1)	
Three or four times	558 (7.8)	504 (90.3)	54 (9.7)		504 (90.3)	54 (9.7)		503 (90.1)	55 (9.9)	
Five or more times	347 (5.1)	321 (92.5)	26 (7.5)		317 (91.4)	30 (8.6)		322 (92.8)	25 (7.2)	
Frequency of lateness for school* (missing = 3722)				<0.001[a]			<0.001[a]			<0.001[a]
Never	3884 (55.7)	2993 (77.1)	891 (22.9)		2952 (76.0)	932 (24.0)		2949 (75.9)	935 (24.1)	
One or two times	2092 (30.0)	1738 (83.1)	354 (16.9)		1708 (81.6)	384 (18.4)		1702 (81.4)	390 (18.6)	
Three or four times	561 (8.1)	506 (90.2)	55 (9.8)		492 (87.7)	69 (12.3)		493 (87.9)	68 (12.1)	
Five or more times	432 (6.2)	391 (90.5)	41 (9.5)		395 (91.4)	37 (8.6)		384 (88.9)	48 (11.1)	
Additional instruction (missing = 4403)				<0.001[a]			<0.001[a]			<0.001[a]
No	5725 (91.1)	4327 (75.6)	1398 (24.4)		4257 (74.4)	1468 (25.6)		4250 (74.2)	1475 (25.8)	
Yes	563 (8.9)	530 (94.1)	33 (5.9)		539 (95.7)	24 (4.3)		537 (95.4)	26 (4.6)	
Learning time (minutes per week; mean, SD)				<0.001[b]			<0.001[b]			<0.001[b]

(*continued*)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	N (%)	N (%)	p-value	N (%)	N (%)	p-value	N (%)	N (%)	p-value
ICT use outside of school for leisure (mean, SD; missing = 3102)	0.06 (1.5)	−0.01 (1.59)	0.34 (1.04)	<0.001[b]	−0.02 (1.6)	0.33 (1.01)	<0.001[b]	0 (1.61)	0.28 (0.98)	<0.001[b]
ICT use outside of school for learning (mean, SD; missing = 3635)	0.2 (1.23)	0.21 (1.3)	0.17 (0.93)	<0.001[b]	0.22 (1.3)	0.14 (0.95)	<0.001[b]	0.22 (1.31)	0.12 (0.89)	<0.001[b]
ICT use at school in general (mean, SD; missing = 3881)	−0.31 (1.19)	−0.28 (1.23)	−0.42 (1.01)	<0.001[b]	−0.28 (1.23)	−0.44 (1.02)	<0.001[b]	−0.26 (1.24)	−0.48 (0.96)	<0.001[b]
Interest in ICT (mean, SD; missing = 4071)	0.15 (1.15)	0.08 (1.2)	0.39 (0.96)	<0.001[b]	0.07 (1.19)	0.41 (0.98)	<0.001[b]	0.07 (1.19)	0.4 (0.98)	<0.001[b]
Perceived ICT competence (mean, SD; missing = 4380)	−0.01 (0.96)	−0.06 (0.98)	0.15 (0.87)	<0.001[b]	−0.06 (0.97)	0.16 (0.91)	<0.001[b]	−0.05 (0.98)	0.13 (0.9)	<0.001[b]
Perceived autonomy related to ICT use (mean, SD; missing = 4358)	−0.02 (1.01)	−0.09 (1.01)	0.24 (0.97)	<0.001[b]	−0.1 (1)	0.25 (1)	<0.001[b]	−0.07 (1.01)	0.16 (0.98)	<0.001[b]
ICT as a topic in social interaction (mean, SD; missing = 4591)	0.22 (0.94)	0.22 (0.96)	0.26 (0.88)	<0.001[b]	0.22 (0.95)	0.25 (0.91)	<0.001[b]	0.24 (0.96)	0.19 (0.88)	<0.001[b]
Information about careers (mean, SD; missing = 2869)	−0.51 (1.04)	−0.54 (1.08)	−0.36 (0.82)	<0.001[b]	−0.55 (1.09)	−0.33 (0.81)	<0.001[b]	−0.55 (1.1)	−0.34 (0.77)	<0.001[b]
Information about the labor market provided by the school (mean, SD; missing = 3109)	−0.41 (0.83)	−0.39 (0.85)	−0.49 (0.72)	<0.001[b]	−0.39 (0.85)	−0.5 (0.71)	<0.001[b]	−0.38 (0.86)	−0.52 (0.68)	<0.001[b]
Information about the labor market provided outside of school (mean, SD; missing = 3109)	0.05 (0.98)	0.05 (1)	0.02 (0.88)	<0.001[b]	0.05 (1.01)	0.04 (0.88)	<0.001[b]	0.05 (1.01)	0.04 (0.86)	<0.001[b]
Non-cognitive skills										
Student's expected occupational status (SEI; mean, SD; missing = 3049)	72.17 (16.67)	71.19 (17.28)	76.67 (12.57)	<0.001[b]	71.17 (17.26)	76.49 (12.97)	<0.001[b]	71.03 (17.27)	76.97 (12.77)	<0.001[b]
Disciplinary climate in test language lessons (mean, SD; missing = 592)	−0.36 (0.99)	−0.41 (0.98)	−0.14 (0.99)	<0.001[b]	−0.41 (0.98)	−0.14 (0.98)	<0.001[b]	−0.41 (0.98)	−0.17 (0.98)	<0.001[b]

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value	*N* (%)	*N* (%)	*p*-value
Teacher support in test language lessons (mean, SD; missing = 601)	0.42 (0.91)	0.41 (0.92)	0.45 (0.87)	<0.001[b]	0.41 (0.92)	0.44 (0.86)	<0.001[b]	0.42 (0.91)	0.43 (0.88)	<0.001[b]
Teacher-directed instruction (mean, SD; missing = 657)	6.41 (23.71)	7.51 (25.53)	0.76 (7.91)	<0.001[b]	7.61 (25.69)	0.51 (6.21)	<0.001[b]	7.55 (25.59)	0.97 (9.2)	<0.001[b]
Perceived feedback (mean, SD; missing = 1005)	−0.18 (0.96)	−0.16 (0.95)	−0.27 (0.98)	<0.001[b]	−0.16 (0.95)	−0.26 (0.99)	<0.001[b]	−0.15 (0.95)	−0.32 (0.98)	<0.001[b]
Parents' emotional support perceived by student (mean, SD; missing = 807)	−0.14 (0.98)	−0.21 (0.98)	0.14 (0.94)	<0.001[b]	−0.21 (0.98)	0.14 (0.95)	<0.001[b]	−0.22 (0.98)	0.16 (0.94)	<0.001[b]
Teacher's stimulation of reading engagement perceived by student (mean, SD; missing = 982)	−0.02 (1.03)	−0.02 (1.03)	0.01 (0.99)	<0.001[b]	−0.02 (1.04)	0 (0.98)	<0.001[b]	−0.02 (1.03)	0 (1)	<0.001[b]
Adaptation of instruction (mean, SD; missing = 944)	−0.09 (1)	−0.14 (1)	0.13 (1)	<0.001[b]	−0.14 (1)	0.13 (1)	<0.001[b]	−0.14 (0.99)	0.1 (1.03)	<0.001[b]
Perceived teacher's interest (mean, SD; missing = 807)	0.2 (0.92)	0.17 (0.91)	0.37 (0.97)	<0.001[b]	0.17 (0.91)	0.35 (0.96)	<0.001[b]	0.17 (0.91)	0.35 (0.96)	<0.001[b]
Joy/like reading (mean, SD; missing = 1036)	0.38 (0.88)	0.33 (0.83)	0.6 (1.02)	<0.001[b]	0.3 (0.82)	0.7 (1.02)	<0.001[b]	0.28 (0.81)	0.8 (1.01)	<0.001[b]
Perception of competitiveness at school (mean, SD; missing = 3948)	0.08 (1.03)	0.04 (1.02)	0.24 (1.03)	<0.001[b]	0.05 (1.02)	0.21 (1.05)	<0.001[b]	0.05 (1.02)	0.2 (1.04)	<0.001[b]
Perception of cooperation at school (mean, SD; missing = 4449)	−0.37 (1.05)	−0.4 (1.06)	−0.23 (1.01)	<0.001[b]	−0.4 (1.07)	−0.26 (1)	<0.001[b]	−0.39 (1.06)	−0.28 (1.02)	<0.001[b]
Attitude toward school learning activities (mean, SD; missing = 1626)	0.33 (0.96)	0.32 (0.98)	0.37 (0.86)	<0.001[b]	0.32 (0.98)	0.37 (0.86)	<0.001[b]	0.32 (0.97)	0.37 (0.87)	<0.001[b]
Competitiveness (mean, SD; missing = 1787)	−0.1 (0.97)	−0.16 (0.97)	0.15 (0.94)	<0.001[b]	−0.16 (0.96)	0.13 (0.96)	<0.001[b]	−0.14 (0.97)	0.07 (0.94)	<0.001[b]

(continued)

Table 1. (continued)

		Proficiency in mathematics			Proficiency in science			Proficiency in reading		
		No	Yes		No	Yes		No	Yes	
	N (%)	N (%)	N (%)	p-value	N (%)	N (%)	p-value	N (%)	N (%)	p-value
Work mastery (mean, SD; missing = 2004)	0.27 (1.01)	0.22 (1.03)	0.45 (0.92)	<0.001[b]	0.22 (1.03)	0.45 (0.92)	<0.001[b]	0.21 (1.03)	0.47 (0.92)	<0.001[b]
General fear of failure (mean, SD; missing = 1993)	0.03 (1.01)	−0.01 (1.01)	0.2 (1)	<0.001[b]	−0.01 (1)	0.21 (1.03)	<0.001[b]	−0.02 (0.99)	0.26 (1.03)	<0.001[b]
Eudaemonia (mean, SD; missing = 2088)	0.09 (0.94)	0.14 (0.92)	−0.12 (1.02)	<0.001[b]	0.15 (0.91)	−0.15 (1.04)	<0.001[b]	0.15 (0.91)	−0.15 (1.04)	<0.001[b]
Positive affect (mean, SD; missing = 2160)	0.06 (1)	0.06 (1.02)	0.07 (0.92)	<0.001[b]	0.07 (1.01)	0.03 (0.96)	<0.001[b]	0.07 (1.01)	0.02 (0.94)	<0.001[b]
Resilience (mean, SD; missing = 2223)	−0.16 (0.97)	−0.19 (0.98)	−0.03 (0.94)	<0.001[b]	−0.18 (0.98)	−0.05 (0.94)	<0.001[b]	−0.17 (0.98)	−0.09 (0.94)	<0.001[b]
Mastery goal orientation (mean, SD; missing = 2170)	0.55 (1.0)	0.53 (1.02)	0.62 (0.89)	<0.001[b]	0.54 (1.02)	0.58 (0.89)	<0.001[b]	0.53 (1.02)	0.62 (0.89)	<0.001[b]
Sense of belonging to school (mean, SD; missing = 2352)	−0.18 (0.97)	−0.23 (0.95)	0 (1.01)	<0.001[b]	−0.23 (0.94)	0 (1.04)	<0.001[b]	−0.23 (0.95)	0.01 (1.03)	<0.001[b]
Student's experience of being bullied (mean, SD; missing = 4297)	0.14 (1.1)	0.17 (1.12)	0.05 (0.99)	<0.001[b]	0.18 (1.13)	0.01 (0.97)	<0.001[b]	0.19 (1.13)	−0.02 (0.95)	<0.001[b]
Metacognition: Understanding and remembering (mean, SD; missing = 2235)	−0.28 (0.99)	−0.43 (0.95)	0.32 (0.91)	<0.001[b]	−0.44 (0.95)	0.32 (0.9)	<0.001[b]	−0.44 (0.95)	0.33 (0.9)	<0.001[b]
Metacognition: Summarizing (mean, SD; missing = 2219)	−0.31 (0.98)	−0.47 (0.95)	0.34 (0.83)	<0.001[b]	−0.49 (0.95)	0.36 (0.79)	<0.001[b]	−0.5 (0.95)	0.38 (0.77)	<0.001[b]
Metacognition: Assessing credibility (mean, SD; missing = 2475)	−0.37 (0.91)	−0.52 (0.85)	0.24 (0.87)	<0.001[b]	−0.54 (0.85)	0.27 (0.85)	<0.001[b]	−0.55 (0.84)	0.29 (0.84)	<0.001[b]

[a]Chi-square test.
[b]ANOVA test.
*In the 2 weeks prior to.
ISCED: International Standard Classification of Education.
ISEI: International Socio-Economic Index of Occupational Status.
ICT: Information and Communication Technology (on a logit scale, where 0 represents the OECD average).

Table 2. Top Proficiency in Mathematics, Science, and Reading Predictors Based on Mean Decrease in Accuracy (MDA), PISA 2018 (*N* = 5345).

	Proficiency in mathematics		Proficiency in science		Proficiency in reading	
Ranking	Predictor	MDA	Predictor	MDA	Predictor	MDA
1	Annual household income	0.103	Metacognition: Assess credibility subscale	0.090	Metacognition: Assess credibility subscale	0.103
2	Metacognition: Summarizing subscale	0.068	Additional instruction	0.086	Additional instruction	0.068
3	Metacognition: Assess credibility subscale	0.064	Metacognition: Summarizing subscale	0.071	Metacognition: Summarizing subscale	0.06
4	Parents' highest occupational status	0.047	Annual household income	0.060	Annual household income	0.047
5	Metacognition: Understanding and remembering subscale	0.045	Metacognition: Understanding and remembering subscale	0.038	Metacognition: Understanding and remembering subscale	0.045
6	Additional instruction	0.038	Parents' highest occupational status	0.036	Parents' highest occupational status	0.038
7	Early childhood education and care	0.037	Perceived autonomy related to ICT use	0.034	Perceived autonomy related to ICT use	0.037
8	Frequency of skipped school classes	0.036	Household possessions	0.032	Household possessions	0.036
9	Grade repetition	0.027	Grade repetition	0.024	Grade repetition	0.027
10	Learning time	0.022	Early childhood education and care	0.022	Early childhood education and care	0.022

ICT: Information and Communication Technology.
Mean decrease in accuracy (MDA) represents how much the predictive performance of a random forest model decreases when the values of a specific feature are randomly permuted.

proficiency in mathematics, science, and reading. Similarly, belonging to the highest quartile of household possessions increased the odds of proficiency in science and reading by 1.48 and 1.91 times, respectively, compared to the first quartile. Early childhood education and care was positively associated with all three outcomes. Additionally, students who repeated a grade and had additional instruction had around 80% and 75% lower odds, respectively, of achieving proficiency in mathematics, science, and reading compared to those who did not repeat.

Among the teaching and learning processes associated with proficiency in mathematics, a one-minute increase per week in learning time was associated with 5% greater odds of achieving this outcome, while the frequency of missed school classes was negatively associated with this outcome. Regarding the scales related to student metacognition and ICT, while the metacognitive scales were positively associated with all three outcomes, perceived autonomy related to ICT use was positively associated with proficiency in science and reading. For each one-point increase on the ICT use scale, the students' odds of proficiency in science increased by 53%, and their odds of proficiency in reading increased by 45%.

According to the log pseudolikelihood, incorporating schools as a random factor in the models resulted in an improvement compared to a usual logistic regression model (*p*-value= <0.001). Furthermore, the results of the intraclass correlation coefficient (ICC) revealed that 31.7%, 26.6%, and 31.9% of the variation in proficiency in mathematics, science, and reading, respectively, were attributable to differences between schools. The area under the receiver operating characteristic

Table 3. Multilevel Logistic Regression of Factors Associated With Proficiency in Mathematics, Science, and Reading, PISA 2018 (Students: $N = 5346$; Schools: $N = 595$).

	Proficiency in mathematics		Proficiency in science		Proficiency in reading	
	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value	OR (95%CI)	*p*-value
Annual household income (Ref.: Up to 10 minimum wages)		<0.001[a]		<0.001[a]		<0.001[a]
More than 10 to 20 minimum wages	3.74 (3.12–4.48)		3.27 (2.73–3.93)		2.83 (2.37–3.38)	
More than 20 minimum wages	4.73 (3.36–6.67)		3.20 (2.22–4.61)		2.36 (1.66–3.36)	
Household possessions (quartile; Ref.: First)				<0.001[a]		<0.001[a]
Second			0.94 (0.69–1.28)		1.28 (0.95–1.71)	
Third			1.07 (0.79–1.45)		1.33 (0.99–1.79)	
Fourth			1.48 (1.07–2.05)		1.91 (1.39–2.62)	
Parents' highest occupational status (ISEI)	1.02 (1.01–1.02)	<0.001[a]	1.01 (1.01–1.02)	<0.001[a]	1.01 (1.00–1.01)	<0.001[a]
Early childhood education and care (Ref.: Less than one year)		0.004[b]		0.003[b]		0.005[b]
Between one and two years	2.51 (1.49–4.24)		1.69 (1.10–2.60)		2.49 (1.38–4.49)	
More than two years	1.94 (1.16–3.23)		1.25 (1.01–1.50)		2.06 (1.16–3.67)	
Grade repetition (yes)	0.20 (0.16–0.25)	<0.001[b]	0.16 (0.13–0.20)	<0.001[b]	0.21 (0.17–0.25)	<0.001[b]
Additional instruction (yes)	0.36 (0.22–0.60)	<0.001[b]	0.35 (0.21–0.59)	<0.001[b]	0.35 (0.21–0.59)	<0.001[b]
Learning time (minute/week)	1.05 (1.03–1.06)	<0.001[b]				
Frequency of skipped school classes (Ref.: Never)		<0.001[a]				
One or two times	0.46 (0.38–0.55)					
Three or four times	0.40 (0.26–0.62)					
Five or more times	0.60 (0.32–1.11)					
Metacognition: Summarizing subscale	1.56 (1.40–1.73)	<0.001[a]	1.58 (1.43–1.74)	<0.001[a]	1.52 (1.38–1.67)	<0.001[a]
Metacognition: Assess credibility subscale	1.95 (1.75–2.18)	<0.001[a]	1.86 (1.67–2.07)	<0.001[a]	1.69 (1.53–1.88)	<0.001[a]
Metacognition: Understanding and remembering subscale	1.43 (1.29–1.59)	<0.001[a]	1.53 (1.39–1.69)	<0.001[a]	1.56 (1.42–1.72)	<0.001[a]
Perceived autonomy related to ICT use			1.53 (1.38–1.70)	<0.001[a]	1.45 (1.32–1.61)	<0.001[a]
ICC (%)	31.7		26.6		31.9	
Log pseudolikelihood	91.61	<0.001	118.26	<0.001	110.12	<0.001

[a]Linear trend test.
[b]Heterogeneity test.
OR: odds ratio.
CI: confidence interval.
ICC: intraclass correlation coefficient.
ICT: Information and Communication Technology (on a logit scale, where 0 represents the OECD average).
ISEI: International Socio-Economic Index of Occupational Status.
The minimum wage in 2018 was equivalent to approximately 299.4 US dollars.

curve (AUC-ROC) for the adjusted multilevel logistic regression models was 92.5% (95% CI: 91.7%–93.3%), 91.7% (95% CI: 90.9%–92.6%), and 92.8% (95% CI: 92.1%–93.5%) for proficiency in mathematics, science, and reading, respectively.

Discussion

To my knowledge, this is the first study that combines machine learning with traditional statistics to identify the most important predictors of proficiency in the PISA mathematics, science, and reading tests. Using an RFA, eight of the top ten predictors identified were common across all three subjects: annual household income, parents' highest occupational status, three metacognition subscales, early childhood education and care, additional instruction, and grade repetition. The remaining predictors varied by subject: frequency of skipped school classes and learning time were specific to mathematics proficiency, while perceived autonomy related to ICT use and household possessions were specific to science and reading proficiency.

My findings identified some key predictors of educational performance observed in countries that consistently score above the OECD average in the PISA. A study that explored the relationship between German students' attitudes toward ICT and their mathematical and scientific literacy, as measured by the 2015 PISA, found that while students' attitudes toward ICT were less important than socioeconomic status, they were more important than ICT autonomy (Lezhnina & Kismihók, 2022). In East Asia, where students consistently excel in mathematics, a study using TIMSS 2019 data from Chinese Taipei, Hong Kong, Japan, Korea, and Singapore identified 11 key predictors of mathematic performance, with students' confidence in mathematics, socioeconomic status, and school emphasis on academic success being the top three (Wang et al., 2023). Both studies, similar to the present study, used the RFA and the mean decrease in accuracy (MDA) metric to rank the importance of predictors.

The relationship between socioeconomic status and school performance is well established in the literature and is reported in PISA reports. However, my analysis suggests that the variables constituting socioeconomic status in the PISA do not have the same level of importance on proficiency in mathematics, science, and reading in Brazil. When ranking the most important predictors, parental occupation status, household possessions, and/or household income were among the top 10 predictors for mathematics, science, and reading performance, and father's and mother's education were not among the top predictors for any of the three outcomes.

It is plausible that parents with higher levels of occupation status are more involved in their children's educational process (Kim et al., 2013). This increased involvement can lead to better relationships with their children's teachers, a higher likelihood of intervening in important school matters, and providing guidance and tutoring for their children's academic performance (Gordon & Cui, 2012), all of which can influence children's performance in mathematics, science, and reading. Additionally, disadvantaged Brazilian students are concentrated in public schools with lower academic performance, while advantaged students attend private schools with better outcomes (IBGE, 2022; Perosa & Dantas, 2017). In Brazilian public schools, unfavorable teaching and learning conditions, such as overworked teachers and overcrowded classrooms, hinder student performance (Martínez García et al., 2021). However, in my findings, household possessions were positively associated with proficiency in mathematics, science, and reading, even after considering the school where the student was enrolled. It is possible that factors associated with family background, not investigated in the present study, influence performance in PISA tests.

Learning time at school was found to be an important predictor for proficiency in mathematics. This finding is consistent with previous research on the importance of learning time in domains as mathematics and reading (Barrios Fernández, 2023; Kidron et al., 2014; Wang et al., 2023). It is likely that students who spend more time studying and practicing are simply exposed to more of the material that they need to master (Kidron et al., 2014). This increased exposure can lead to better understanding, retention, and application of the material. Moreover, learning time can also provide students with opportunities to receive feedback and guidance from teachers and peers. This feedback can help students identify their strengths and weaknesses and can also provide them with strategies for improvement (Aronson et al., 1999). Nonetheless, learning time can also create opportunities for students to develop self-discipline and perseverance. These qualities can impact performance in mathematics achievement (Barrios Fernández, 2023).

The culture of grade repetition is strong in Brazilian society, extending beyond school walls (Oliveira, 2016). A noticeable proportion of teachers and parents in Brazil still believe that grade retention can lead to academic recovery for students who fail to meet desired academic levels at the end of a grade (Oliveira, 2016). However, grade repetition was one of the top 10 predictors of mathematics, science, and reading performance in the present study and was negatively associated with these outcomes. Similar findings also were observed in previous studies. Wang et al.'s meta-analysis revealed a negative association between school repetition and mathematics achievement in the PISA (Wang et al., 2023). Additionally, grade repetition accounted for 10% of the variation in reading performance in the 2018 PISA assessment across included countries (OECD, 2020). Failure means an extension of the time spent at school by one year. This means that students lose their classmates and have to join a new group of students who are younger than them (Gottfredson et al., 1994). Thus, student retention can harm low-performing students by limiting their learning opportunities and motivation (Goos et al., 2021). Additionally, an exploratory analysis of the PISA data in Brazil showed that students who failed tend to change schools more frequently, likely seeking schools with less demanding scholar expectations.

My findings suggested that the number of years a student spent in early childhood education and care was one of the top 10 predictors of proficiency in mathematics, science, and reading and is positively associated with these outcomes. Early childhood education and care is widely believed to be beneficial for children's development. When preschool is taken into account, research presents a number of benefits for children who attend this stage of education, such as the reduction of mortality in this age range, higher cognitive development, more time spent at school, reduction of school year repetition and school dropout, and even more vocabulary acquisition due to very early exposure to different environments (Belsky et al., 2007). Additionally, a study conducted in Brazil showed that children who attend preschool have, on average, one more year of formal education than those who entered primary school directly and a 32% higher chance of finishing secondary school (Becker, 2007).

Finally, my analysis revealed that adolescent students' perceived ICT autonomy was one of the top 10 predictors of proficiency in science and reading. Perceived autonomy related to ICT usage refers to one's perceptions of personal independence (i.e., lack of external constraints or controls) in competently using digital media and digital devices, including desktop computers, portable laptops, notebooks, smartphones, tablet computers, cell phones without Internet access, game consoles, and Internet-connected television. It was positively associated with these outcomes in the present study and in previous studies using PISA 2009, 2012, 2015, and 2018 data (Areepattamannil & Santos, 2019). Additionally, Courtney et al. (2022) found that students who perceived themselves as autonomous and competent in ICT use develop positive views and feelings toward science, such as self-efficacy, enjoyment, and interest in science. These findings corroborate the assumptions of the self-determination theory and the ICT engagement concept (Ryan & Deci, 2017), suggesting that academically successful students exhibit a heightened sense of self-direction and control in ICT-related activities (autonomy). Given these relationships, it is plausible that a cluster of student attributes exists that is associated with positive beliefs and attitudes toward learning in ICT and in general.

Strengths and Limitations

The present study has several strengths. It is based on the PISA 2018 database, which encompasses a wide range of variables, including socioeconomic background, individual characteristics, and educational experiences, enabling comprehensive analyses of the factors influencing student achievement. Additionally, the large sample size of Brazilian students in the PISA database enhances the statistical power and generalizability of findings. Moreover, random forest, a machine learning algorithm frequently employed to identify the most effective predictors in a dataset, was utilized. One of the primary benefits of using random forest is its high accuracy and ability to handle large and intricate datasets. The out-of-bag accuracy exceeding 80% in the random forest analysis models developed in this study indicates that the models are likely to generalize well to new data.

Combining machine learning for variable classification with traditional statistics leverages the strengths of both methods: machine learning's ability to handle complex relationships and traditional statistics's high interpretability for understanding the effects of those variables. The RFA was chosen for this study due to its suitability for large datasets, its ability to handle multiple predictors, and its effectiveness with highly correlated variables—characteristics found in the PISA dataset. Additionally, its widespread use in studies combining machine learning and traditional statistics further supports this choice (Lezhnina & Kismihók, 2022; Wang et al., 2023). Other approaches, such as penalized regression methods, can also be used. A recent study using PISA data to evaluate predictors for anxiety suggested that penalized regression methods like Least Absolute Shrinkage and Selection Operator (LASSO) and Enet may outperform random forest in predictor classification tasks. However, the difference in mean root-mean-square deviation between the methods was minimal. Additionally, anxiety, as an outcome, may be less directly influenced by ecological factors like school environment compared to proficiency levels in PISA tests (Immekus et al., 2022). Furthermore, although these approaches are capable of handling highly correlated variables, they still present limitations in such situations (Herawati et al., 2018).

It is also important to consider that the data collected impose some limitations on the analyses. First, the cross-sectional design of this study limits analysis of temporal associations. For instance, additional instruction was negatively associated with proficiency. Given the cross-sectional design of our study, it is likely that those who perform the poorest are those who seek the most help outside of school. Additionally, it is also plausible that low performance precedes grade repetition (Fischbach et al., 2013). Second, the instruments used in the study were self-report instruments. Although self-report instruments have been extensively employed by psychologists and educators for decades, their utilization in research is susceptible to response bias. This includes the tendency to endorse items positively regardless of content or to endorse items based on the social desirability of the response. Third, the substantial number of missing responses for some variables, even when using chained equations to impute complete datasets, can introduce bias into the results. This stems from the fact that the imputed values are based on data from participants who completed the questionnaire, and this data may not be representative of the total population. Finally, although the restricted number of predictors in logistic regression reduces the likelihood of complex interactions, they cannot be entirely ruled out.

While dichotomizing proficiency levels into "Low Proficient" and "Proficient" simplifies the interpretation of results and helps in formulating targeted educational policies, some information may be lost by combining levels. However, this is likely minimal in our sample due to the distribution of proficiency levels. The vast majority of Brazilian students (74.1% in reading, 88.0% in mathematics, and 81.8% in science) scored at levels 1 or 2, with a very small percentage (1%–2%) reaching the advanced levels of 5 or above. This distribution strengthens the dichotomization approach, allowing for a clearer focus on the primary challenge of raising baseline proficiency levels in Brazil.

This study focuses on individual predictors, yet approximately 30% of the variation in mathematics, science, and reading proficiency was attributable to differences between schools. In Brazil, inequalities in educational access and a preference for attending schools close to home may contribute to this variation, as schools tend to reflect homogenous socioeconomic groups (Lépine, 2015). However, the sample in this study had a limited number of students per school, with a maximum of 30. Although the high number of schools and low number of students per school allows the use of single-level random forests with results similar to multilevel random forests, decreasing the risk of overfitting, the low internal variability within these smaller groups may have led to an overestimation of the variation between schools in multilevel logistic regression analyses (Mangino & Finch, 2021). Finally, this study did not incorporate classroom-level grouping due to the limited number of students per classroom. The PISA offers school-level variables, making it possible for future studies to address these school-based influences in more detail.

Future Research Directions

Large-scale school performance assessments are essential tools for effective planning of education systems. However, the abundance of information they generate necessitates approaches that facilitate a deeper understanding of the data. Implementing an RFA to select variables for inclusion in traditional statistical analyses can enhance data management and simplify result interpretation. A relatively simple RFA implementation was chosen to align with the data distribution, ensure greater comprehensibility, and facilitate reproducibility with a Python script. Nevertheless, future research can explore more sophisticated algorithms for the classification task (Romero & Ventura, 2020; Ryo & Rillig, 2017).

Conclusion

The present study used the RFA to highlight the 10 most important predictors for proficiency in each PISA test and measured their effects using traditional statistical methods.

The top predictors were highly similar for mathematics, science, and reading tests. The results point to early childhood education and care, grade repetition, additional instruction, metacognition, frequency of skipped school classes, learning time, and perceived autonomy related to ICT use as key factors for high performance. These findings urge Brazilian policymakers and educators to prioritize initiatives that strengthen early childhood programs, minimize grade repetition, promote effective learning strategies, and empower students with ICT.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Pedro San Martin Soares https://orcid.org/0000-0001-8974-7503

Supplemental Material

Supplemental material for this article is available online.

References

Afshartous, D., & de Leeuw, J. (2005). Prediction in multilevel models. *Journal of Educational and Behavioral Statistics*, *30*(2), 109–139. A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association. https://doi.org/10.3102/10769986030002109

Aksu, G., & Güzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using decision-tree method: Turkey sampling. *Dialysis, Transplantation & Burn: Medical Journal of the Turkish Transplantation and Burn Foundation*, *41*(185), 101–121. https://doi.org/10.15390/eb.2016.4766

Areepattamannil, S., & Santos, I. M. (2019). Adolescent students' perceived information and communication technology (ICT) competence and autonomy: Examining links to dispositions toward science in 42 countries. *Computers in Human Behavior*, *98*(5), 50–58. https://doi.org/10.1016/j.chb.2019.04.005

Aronson, J. Z., Zimmerman, J., & Carlos, L. (1999). *Improving student achievement by extending school: Is it just a matter of time?* WestEd.

Barrios Fernández, A. (2023). *Instruction time and educational outcomes*. IZA World Labor. https://doi.org/10.15185/izawol.509

Becker, F. (2007). Early Childhood Education in Brazil: The obstacles to a successful experience. *Revista de La Sede Academica de Mexico de La Facultad Latinoamericana de Ciencias Sociales*, *5*(2), 515–537.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T., NICHD Early Child Care Research Network. (2007). Are there long-term effects of early child care? *Child Development*, *78*(2), 681–701. https://doi.org/10.1111/j.1467-8624.2007.01021.x

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

Breiman, L. (2001b). The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, *15*(4), 233–234. https://doi.org/10.1038/nmeth.4642

Cheung, K.-C., Sit, P.-S., Soh, K.-C., Ieong, M.-K., & Mak, S.-K. (2014). Predicting academic resilience with reading engagement and demographic variables: Comparing Shanghai, Hong Kong, Korea, and Singapore from the PISA perspective. *The Asia-Pacific Education Researcher*, *23*(4), 895–909. https://doi.org/10.1007/s40299-013-0143-4

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1), 270. https://doi.org/10.1186/s12859-018-2264-5

Courtney, M., Karakus, M., Ersozlu, Z., & Nurumov, K. (2022). The influence of ICT use and related attitudes on students' math and science performance: Multilevel analyses of the last decade's PISA surveys. *Large-Scale Assessments in Education*, *10*(1), 1–26. https://doi.org/10.1186/s40536-022-00128-6

Fischbach, A., Keller, U., Preckel, F., & Brunner, M. (2013). PISA proficiency scores predict educational outcomes. *Learning and Individual Differences*, *24*(6), 63–72. https://doi.org/10.1016/j.lindif.2012.10.012

Friedman, J. H. (1997). Data mining and statistics What's the connection? Computing science and statistics: Mining and modeling massive data sets in science, engineering, and business [Symposium]. In *29th Symposium on the Interface*.

Gamazo, A., & Martínez-Abad, F. (2020). An exploration of factors linked to academic performance in PISA 2018 through data mining techniques. *Frontiers in Psychology*, *11*(1), 575167. https://doi.org/10.3389/fpsyg.2020.575167

Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., & Jose Rodriguez-Conde, M. (2018). Assessment of factors related to school effectiveness in PISA 2015. A multilevel analysis. *Revista de La Sociedad Española de Farmacia Hospitalaria*, *379*(3), 56–78. https://doi.org/10.4438/1988-592X-RE-2017-379-369

Gómez-Fernández, N., & Mediavilla, M. (2021). Exploring the relationship between information and communication technologies (ICT) and academic performance: A multilevel analysis for Spain. *Socio-Economic Planning Sciences*, *77*(5), 101009. https://doi.org/10.1016/j.seps.2021.101009

Goos, M., Pipa, J., & Peixoto, F. (2021). Effectiveness of grade retention: A systematic review and meta-analysis. *Educational Research Review*, *34*(3), 100401. https://doi.org/10.1016/j.edurev.2021.100401

Gordon, M. S., & Cui, M. (2012). The effect of school-specific parenting processes on academic achievement in adolescence and young adulthood. *Family Relations*, *61*(5), 728–741. https://doi.org/10.1111/j.1741-3729.2012.00733.x

Gottfredson, D. C., Fink, C. M., & Graham, N. (1994). Grade retention and problem behavior. *American Educational Research Journal*, *31*(4), 761–784. https://doi.org/10.3102/00028312031004761

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In *2015 international symposium on educational Technology (ISET)* (pp. 125–128). https://doi.org/10.1109/ISET.2015.33

Herawati, N., Nisa, K., Setiawan, E., Nusyirwan, N., & Tiryono, T. (2018). Regularized multiple regression methods to deal with severe multicollinearity. *International Journal of Statistics and Applications*, *8*(4), 167–172. https://doi.org/10.5923/j.statistics.20180804.02

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*, *659*(1), 48–62. https://doi.org/10.1177/0002716215570279

IBGE. (2022). *Synthesis of social indicators: An analysis of the living conditions of the Brazilian population*. IBGE.

Immekus, J. C., Jeong, T.-S., & Yoo, J. E. (2022). Machine learning procedures for predictor variable selection for schoolwork-related anxiety: Evidence from PISA 2015 mathematics, reading, and science assessments. *Large-Scale Assessments in Education*, *10*(1), 30. https://doi.org/10.1186/s40536-022-00150-8

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS One*, *13*(8), Article e0201904. https://doi.org/10.1371/journal.pone.0201904

Karakolidis, A., Pitsia, V., & Emvalotis, A. (2016). Examining students' achievement in mathematics: A multilevel analysis of the Programme for international student assessment (PISA) 2012 data for Greece. *International Journal of Educational Research*, *79*(5), 106–115. https://doi.org/10.1016/j.ijer.2016.05.013

Kell, M., & Kell, P. (2014). Global testing: PISA, TIMSS and PIRLS. In M. Kell & P. Kell (Eds.), *Literacy and language in East Asia: Shifting meanings, values and approaches* (pp. 33–49): Springer. https://doi.org/10.1007/978-981-4451-30-7_5

Kidron, Y., & Lindsay, J. (2014). *The effects of increased learning time on student academic and non-academic outcomes: Findings from a meta-analytic review*. Institute of Education Sciences.

Kim, Y., Sherraden, M., & Clancy, M. (2013). Do mothers' educational expectations differ by race and ethnicity, or socioeconomic status? *Economics of Education Review*, *33*(2), 82–94. https://doi.org/10.1016/j.econedurev.2012.09.007

Knighton, T. (2006). *Educational outcomes at age 19 associated with reading ability at age 15*. Statistics Canada.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (2016). *Assessing contexts of learning: An international perspective*. Springer International Publishing. https://doi.org/10.1007/978-3-319-45357-6

Lazarevic, L. B., & Orlic, A. (2018). PISA 2012 mathematics literacy in Serbia: A multilevel analysis of students and schools. *Psihologija*, *51*(4), 413–432. https://doi.org/10.2298/psi170817017l

Lépine, A. (2015). *School reputation and school choice in Brazil: A regression discontinuity design (No. 38)*. Department of Economics- FEA/USP.

Lezhnina, O., & Kismihók, G. (2022). Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. *International Journal of Research and Method in Education*, *45*(2), 180–199. https://doi.org/10.1080/1743727X.2021.1963226

Mangino, A. A., & Finch, W. H. (2021). Prediction with mixed effects models: A Monte Carlo simulation study. *Educational and Psychological Measurement*, *81*(6), 1118–1142. https://doi.org/10.1177/0013164421992818

Martínez-Abad, F., Gamazo, A., & Rodríguez-Conde, M.-J. (2020). Educational Data Mining: Identification of factors associated with school effectiveness in PISA assessment. *Studies In Educational Evaluation*, *66*(3), 100875. https://doi.org/10.1016/j.stueduc.2020.100875

Martínez García, J. S., Oinonen, E., Merino, R., & Perosa, G. (2021). Education and inequality in Finland, Spain and Brazil. In *Towards a comparative analysis of social inequalities between Europe and Latin America*. https://doi.org/10.1007/978-3-030-48442-2_4

Mostafa, J., & Silva, K. (2007). *Brazil's single registry experience: A tool for pro-poor social policies*. International Labour Organization.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons B*, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

OECD. (2007). *PISA 2006 science competencies for tomorrow's world: Volume 1: Analysis: Science competencies for tomorrow's world: Volume 1: Analysis*. OECD Publishing. https://doi.org/10.1787/9789264040014-en

OECD. (2019a). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OECD. (2019b). *PISA 2018 results (Volume I) what students know and can do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD. (2019c). *PISA 2018 results (Volume II) where all students can succeed: Where all students can succeed*. OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2020). *PISA 2018 results (Volume V) effective policies, successful schools*. OECD Publishing. https://doi.org/10.1787/ca768d40-en

Oliveira, A. S. (2016). *Progressão continuada e outros dispositivos escolares: êxito e fracasso escolar nos anos iniciais do ensino fundamental*. [Universidade de São Paulo]. https://doi.org/10.11606/t.48.2014.tde-17122014-094918

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. arXiv. http://arxiv.org/abs/1201.0490

Peng, Y., Wang, Y., & Hu, J. (2023). Examining ICT attitudes, use and support in blended learning settings for students' reading performance: Approaches of artificial intelligence and multilevel model. *Computers & Education*, *203*(12), 104846. https://doi.org/10.1016/j.compedu.2023.104846

Perosa, G. S., & Dantas, A. S. R. (2017). The choice of private schools in families from low-income groups. *Educação e Pesquisa*, *43*(4). https://doi.org/10.1590/S1517-9702201704177976

Pholphirul, P. (2017). Pre-primary education and long-term education performance: Evidence from Programme for international student assessment (PISA) Thailand. *Journal of Early Childhood Research*, *15*(4), 410–432. https://doi.org/10.1177/1476718X15616834

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355

Royal Society. (2017). *Machine learning: The power and Promise of computers that learn by example*. Royal Society.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications.

Ryo, M., & Rillig, M. C. (2017). Statistically reinforced machine learning for nonlinear patterns and variable interactions. *Ecosphere*, *8*(11), Article e01976. https://doi.org/10.1002/ecs2.1976

Schleicher, A. (2007). Can competencies assessed by PISA be considered the fundamental school knowledge 15-year-olds should possess? *Journal of Educational Change*, *8*(4), 349–357. https://doi.org/10.1007/s10833-007-9042-x

She, H.-C., Lin, H.-S., & Huang, L.-Y. (2019). Reflections on and implications of the Programme for International Student Assessment 2015 (PISA 2015) performance of students in Taiwan: The role of epistemic beliefs about science in scientific literacy. *Journal of Research in Science Teaching*, *56*(10), 1309–1340. https://doi.org/10.1002/tea.21553

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education Review*, *44*(4), 387–407. https://doi.org/10.1080/03050060802481413

Touron, J., Lopez-Gonzalez, E., Lizasoain Hernandez, L., Garcia San Pedro, M. J., & Navarro Asencio, E. (2018). Spanish high and low achievers in science in PISA 2015: Impact analysis of some contextual variables. *Revista de La Sociedad Española de Farmacia Hospitalaria*, *380*(2), 156–184. https://doi.org/10.4438/1988-592X-RE-2017-380-376

Unesco. (1997). *International standard classification of education: Isced 1997*. Unesco.

Wagner, P., & Merlo, J. (2014). Discriminatory accuracy of a random effect in multilevel logistic regression. *International Journal of Epidemiology*, *44*(5). https://doi.org/10.1093/IJE/DYV097.190

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in PISA: A systematic review. *Large-Scale Assessments in Education*, *11*(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Williamson, B. D., Gilbert, P. B., Carone, M., & Simon, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, *77*(1), 9–22. https://doi.org/10.1111/biom.13392

Wright, M. N., & König, I. R. (2019). Splitting on categorical predictors in random forests. *PeerJ*, *7*(1), Article e6339. https://doi.org/10.7717/peerj.6339

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393