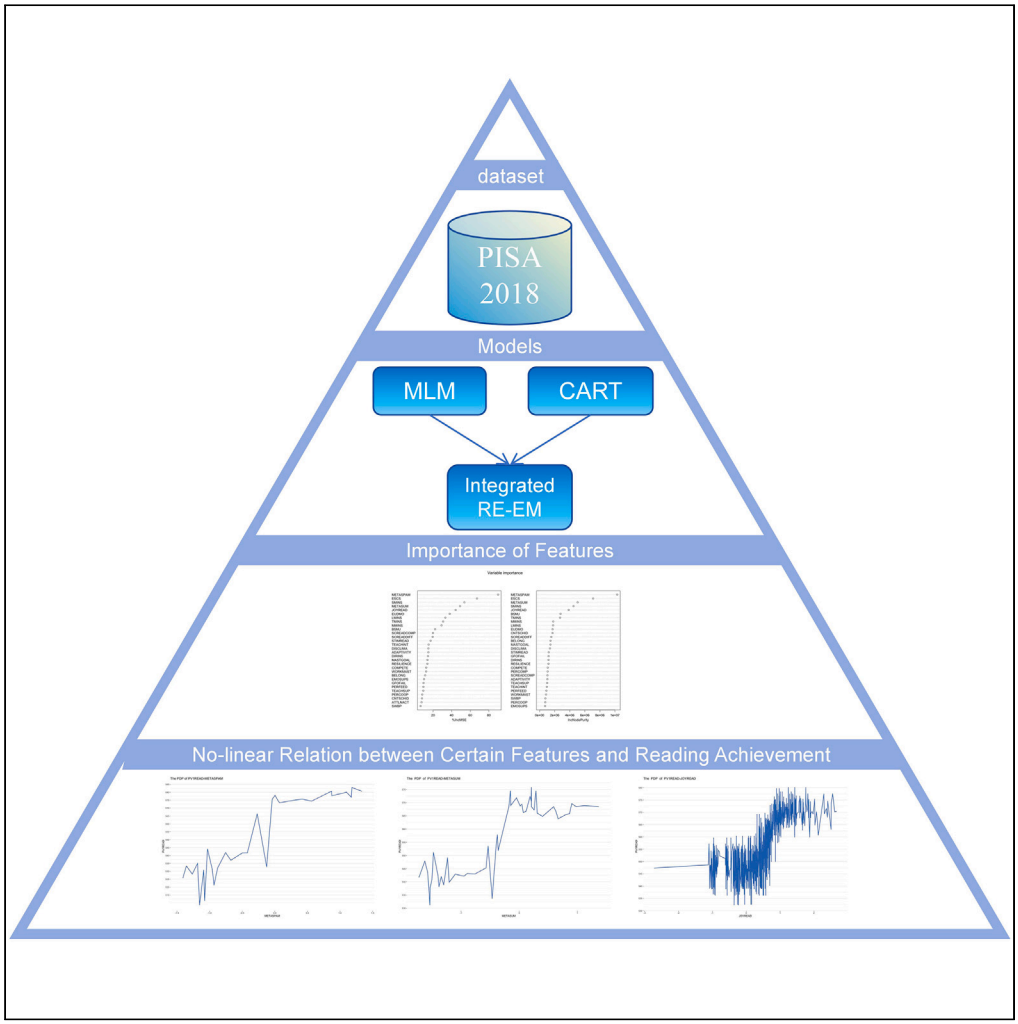
iScience

CellPress
OPEN ACCESS

Article

Identifying key factors of reading achievement: A machine learning approach



Hao Liu, Dongxia Yang, Shangran Nie, Xi Chen

liuhao@bnu.edu.cn

Highlights

The RE-EM regression tree was first introduced for education nested data

Several key factors for reading achievement were identified

A nonlinear impact of key factors on reading achievement was discovered

Liu et al., iScience 27, 110848
October 18, 2024
https://doi.org/10.1016/j.isci.2024.110848

iScience

CellPress
OPEN ACCESS

Article

Identifying key factors of reading achievement: A machine learning approach

Hao Liu,[1,4,*] Dongxia Yang,[1] Shangran Nie,[2] and Xi Chen[3]

SUMMARY

This article explored the influencing factors of digital reading achievement based on the PISA 2018 assessment of students' reading achievement. An integrated Random Effect-Expectation Maximization (RE-EM) regression tree model was the first constructed to address the shortcomings of traditional machine learning methods for nested data estimation and the limitations of traditional linear models in handling complex data. Our study identified the key variables for the feature selection in the integrated RE-EM regression tree model include various aspects of Meta-cognition, as well as the affective element of Joy/Liking for Reading. Notably, this study found that Meta-cognition: Assess Credibility exhibits a ceiling effect on reading achievement, where the marginal effect on reading achievement significantly diminishes at the higher variable values. Additionally, Meta-cognition: Summarizing and Joy/Liking for Reading both demonstrate an approximately S-shaped curve influence on reading achievement. These findings were discussed in critical theoretical and policy implications.

INTRODUCTION

With the in-depth development of the knowledge economy and the development of lifelong learning, reading is one of the most important ways for people to acquire information and adapt to life in terms of academic achievement and practical application. In recent years, the significance of reading has been recognized globally, with UNESCO and various nations, including those in Europe and the United States, emphasizing it as a key lifelong learning skill. The inclusion of digital reading in the 2018 Program for International Student Assessment (PISA) by the OECD underscores this trend. Moreover, public libraries play a vital role in fostering a reading culture, as evidenced by initiatives such as China's 14th Five-Year Plan, which focuses on promoting and facilitating reading among the public.

PISA, as the most extensive global initiative for assessing and monitoring educational quality, has been instrumental in guiding countries to track educational standards, reform educational practices, refine policies, and elevate educational levels. Its assessment outcomes and the factors influencing them have been the subject of extensive research and discussion. PISA 2018 defined reading achievement as understanding, using, evaluating, reflecting on, and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society.[1] From the definition provided by PISA, it is evident that reading achievement nowadays is no longer a skill acquired solely in the early stages of education but rather an evolving set of skills and strategies. The focus has shifted from mere collection and memorization to the acquisition and utilization of information.[2]

Common methodologies in this field include the Hierarchical Linear Model (HLM) and various machine learning models. While HLM is apt for nested data analysis, its limitation lies in handling a restricted set of variables, thus hindering the integration of multiple variables into a comprehensive framework.[3] On the other hand, machine learning models, such as the Classification and Regression Trees (CART), adopt a top-down recursive division of datasets into distinct feature spaces. However, their effectiveness in fitting nested data is often suboptimal, leading to weaker accuracy in model construction and a tendency toward greater bias.

Given the complexity and extensive range of big data in education, it is imperative to identify scientific methodologies for examining its underlying factors. This study leverages data from the PISA 2018, focusing on student reading achievement, to introduce a novel machine learning algorithm. This study integrates a hierarchical approach into an existing machine learning framework, creating the integrated Random Effect-Expectation Maximization (RE-EM) Regression Tree Model. This model is then evaluated against traditional algorithms, assessing its strengths and limitations. Utilizing this optimized machine learning approach, the study constructs a predictive model for reading achievement across four Chinese provinces/municipalities. This study serves two primary objectives: firstly, to employ a novel approach that integrates multilevel modeling with machine learning models, and to use interpretable machine learning techniques to demystify the "black box" models. This combined approach is more suitable for handling the nested, large-scale data of PISA than using multilevel modeling or

[1]Collaborative Innovation Centre for Assessment of Basic Education Quality, Beijing Normal University, Beijing, China
[2]Business School, The University of Sydney, Camperdown, NSW, Australia
[3]School of Social Science, Tsinghua University, Beijing, China
[4]Lead contact
*Correspondence: liuhao@bnu.edu.cn
https://doi.org/10.1016/j.isci.2024.110848

machine learning models in isolation; secondly, to discover significant factors and mechanisms influencing students' digital reading achievement in China, thereby offering informed implications to enhance reading outcomes and overall educational quality.

Literature review

Understanding reading achievement based on the tri-dimensional theoretical framework

Bronfenbrenner posited that an individual's physiological and psychological states are influenced by interconnected and multi-layered environmental systems extending from the innermost to the outermost layers.[4] According to the Ecological Systems Theory, we constructed an analytical framework to examine the factors influencing students' reading achievement, which encompasses three dimensions: the individual level, the school level, and the family level.

The primary motivation for realizing self-worth is found in self-drive, with personal factors playing a significant role in students' reading achievement. Research by Hu and Wang identified key factors affecting reading comprehension from an individual-learner perspective, including cognitive levels, emotional differences, and reading strategies.[5] Similarly, Liao and Wang highlighted the impact of cognitive and emotional factors, along with emotional support from significant others, on reading fluency.[6] Logan, Medford and Hughes argued that intrinsic motivation explains the differences in the enhancement of reading skills among students with lower achievement levels.[7] Chen explored the influence of innate and self-motivation on reading proficiency, finding a positive correlation with factors such as metacognitive reading strategies, self-educational expectations, and reading interest.[8] Schoor observed that learners' self-efficacy and the intrinsic value of the task significantly enhance reading fluency, mediated by reading style and behavior.[9] Nalipay, King and Cai emphasized the importance of learner relationships, autonomy, and competitiveness, as outlined in self-determination theory, in relation to reading fluency, noting a clear effect of these individual psychological factors on reading achievement.[10]

School factors have also been extensively researched, generally categorized into two major aspects: soft power, encompassing teacher strength, curriculum arrangement, and school influence; and hard power, including physical infrastructure and campus construction. Liu and Kang used PISA 2018 data from four Chinese provinces/municipalities, identifying those factors such as school location and size, class size, teacher-to-student ratio, and student-to-computer ratio significantly predict student achievement, with notable interactions between school type, class size, and Socioeconomic and Cultural Status (ESCS).[11] Berkowitz pointed out that a positive social atmosphere in schools can mitigate the strength of the association between Socioeconomic Status (SES) and academic performance, thereby narrowing the literacy achievement gap among students from different economic backgrounds.[12] Jia and Zhang discussed how varying teaching styles impact learners' reading achievement, with cognitive activation strategies and teacher-adapted instruction positively influencing achievement, in contrast to the negative effect of teacher-directed instructional strategies.[13] Teacher support can foster positive emotions and mitigate negative ones, consequently enhancing students' academic engagement and enjoyment.[14] Teacher can enhance students' reading comprehension performance by employing appropriate teaching styles, such as providing effective and challenging learning tasks and using motivational teaching strategies to stimulate students' interest in reading.[15] Ning et al. found that a disciplined school classroom environment substantially boosts students' self-expectations and motivation, thereby aiding in the enhancement of reading achievement.[16]

Furthermore, an increasing body of research suggests a pivotal role for family factors in developing reading skills. Konstantopoulos and Borman even concluded that family engagement is a better predictor of student achievement than school engagement.[17] Banerjee noted the challenge of compensating for disparities in home education within the school setting.[18] Home reading environment,[19,20] parental reading support,[20,21] and parental beliefs and values about reading[22] have been proven to effectively influence students' reading achievement. In addition, socioeconomic and cultural status[23] and parents' education level[24,25] are key predictors of students' reading achievement. Families with better economic conditions possess greater cultural capital, which enables them to provide more reading resources and educational opportunities to enhance their children's reading skill.[26] Additionally, there is a strong positive correlation between parents' education level and the time they invest in their children,[27] which indirectly influences children's academic performance through parental involvement.[28] Netten, Voeten, Droop, and Verhoeven identified socioeconomic and cultural status as crucial predictors of students' reading abilities,[23] while the level of parents' education has been similarly highlighted.[24,25] Families with better economic conditions possess more cultural capital, which facilitates access to reading resources and educational opportunities, thereby enhancing reading capabilities.[26] Additionally, a strong positive correlation exists between parents' educational levels and the amount of time they spend with their children,[27] which in turn indirectly affects children's academic performance through parental involvement.[28]

Methods for analyzing nested data

In educational research, the prevalence of nested data structures is a notable characteristic. This field encompasses various educational forms, including family, school, and social education. Most existing statistical studies on educational data predominantly utilize traditional statistical methods such as structural equation modeling (SEM) and multilevel linear modeling (MLM) to examine the relationships between variables. For instance, Huang and Benoliel used SEM based on Singapore's PISA data to investigate whether principals' time allocation can influence student performance by shaping the school climate.[29] Wu and Zhang employed MLM using data from six countries (regions) in PISA to study the impact of individual attributes, family background, and school characteristics on students' global competencies.[30] Although these methods have yielded effective results in empirical research, their limitations are evident. For example, the models mentioned above are linear and assume no mutual influence among independent variables, making them unable to identify nonlinear relationships and complex interactions between variables. When nonlinear relationships exist, linear models may overestimate or underestimate the strength of relationships. Additionally, these models require certain assumptions to be met for valid results, such as normality and homoscedasticity. However,

Table 1. Explanation of relevant variables at the student level

Level	Factor	Variable	Description
Student	Career Expectations	BSMJ	Student's Expected Occupational Status
	Meta-cognition Strategies	UNDREM	Meta-cognition: Understanding and Remembering
		METASUM	Meta-cognition: Summarizing
		METASPAM	Meta-cognition: Assess Credibility
	Self-Concept of Reading	JOYREAD	Joy/Like Reading
		SCREADCOMP	Self-Concept of Reading: Perception of Competence
		SCREADDIFF	Self-Concept of Reading: Perception of Difficulty
	Personality variables	COMPETE	Competitiveness (WLE)
		WORKMAST	Work Mastery (WLE)
		GFOFAIL	General Fear of Failure (WLE)
		RESILIENCE	Resilience (WLE)
		MASTGOAL	Mastery Goal Orientation (WLE)
	Student well-being	EUDMO	Eudaemonia: Meaning in Life (WLE)
		SWBP	Subjective Well-being: Positive Affect (WLE)

many real-world scenarios do not necessarily satisfy these assumptions, and in the context of big data, these models often exhibit poor applicability and robustness.

The integration of linear models with machine learning techniques has been a focus for many researchers. Research by Lin and Luo investigated the M-CART method, a novel algorithm blending multinomial-logit (M-logit) and single-CART (S-CART) within an expectation maximization framework. Their findings indicated that M-CART significantly enhances classification accuracy, sensitivity, and specificity in multilevel data modeling compared to traditional approaches such as M-logit, S-CART, and single-logistic regression.[31] Sela and Simonoff initially introduced the RE-EM regression tree model, combining longitudinal and clustered mixed-effect modeling structures with tree-based estimation methods' flexibility.[32] Subsequently, Fu and Simonoff developed an unbiased RE-EM algorithm, improving upon Sela and Simonoff's original method, which revised RE-EM regression tree model demonstrated unbiasedness, enhanced prediction accuracy, and more accurate tree structure recovery.[33] Li (2019) applied the RE-EM regression tree model to structured medical system data, revealing its efficacy in identifying critical relationships among predictor variables in nested data, thereby improving model fit.[3]

The present investigation advances the field by addressing the methodological limitations inherent in the analysis of nested data through the implementation of the Random Effect-Expectation Maximization (RE-EM) regression tree model. This innovative approach is poised to enrich the research landscape concerning the determinants of reading achievement. The basic machine learning architecture of the RE-EM model is predicated on a decision tree algorithm. However, given the expansive dataset and variable complexity inherent in the Program for International Student Assessment (PISA), reliance on a solitary decision tree model may yield substantial predictive inaccuracies, thereby undermining both the precision and scientific rigor of the findings. To circumvent these limitations, this study introduces a novel methodological contribution by embedding the RE-EM regression tree model within a random forest algorithm. This integration not only enhances the robustness of the model but also leverages the inherent strengths of the RE-EM framework. Consequently, this amalgamated model is anticipated to exhibit markedly improved predictive capabilities. The practical application of this refined machine learning methodology is demonstrated through the development of a predictive model for reading achievement across four select provinces/municipalities in China. This model aims to systematically dissect and elucidate the multifaceted factors influencing digital reading proficiency, thereby contributing significantly to the academic discourse in this domain.

Methods

Data sources and selection of variables

The source of data for this article is from Program for International Student Assessment (PISA 2018), which was chosen to be analyzed from a total of 361 schools and 12,058 students in four provinces/municipalities in China: Beijing, Shanghai, Jiangsu, and Zhejiang.

In an effort to provide a nuanced representation of each student's reading achievement level, the Program for International Student Assessment (PISA) generates ten plausible values (PVs), each randomly drawn from the distribution of proficiency in reading achievement. The OECD[1] posits that for extensive datasets, the error introduced by utilizing a singular PV is insubstantial, thus ensuring the reliability of the resultant data. In alignment with this approach, the current study adopts PV1 as the dependent variable, reflecting students' reading achievement in congruence with the output domain of the machine learning algorithm. Grounded in a comprehensive literature review, 35 predictor variables have been meticulously selected, corresponding to the input domain of the machine learning framework. These variables, encapsulating diverse dimensions pertaining to students, families, and schools, are systematically detailed in Tables 1 and 2. To address the challenge of missing data, this study employs a school-based grouping strategy, applying the median imputation technique

Table 2. Explanation of relevant variables at the family and school level

Level	Factor	Variable	Description
Family	Variables	DURECEC	Duration of Early Children Education and Care
		EMOSUPS	Parents' Emotional Support Perceived by Student
		ESCS	Socioeconomic and Cultural Status
School	Study time	MMINS	Math Learning Time
		LMINS	Reading Learning Time
		SMINS	Science Learning Time
		TMINS	Total Learning Time
	Reading course	DISCLIMA	Disciplinary Climate in Test Language Lessons
		TEACHSUP	Teacher Support in Test Language Lessons
		DIRINS	Teacher-Directed Instruction
		PERFEED	Perceived Feedback
		STIMREAD	Teacher's Stimulation of Reading Engagement Perceived by Student
		ADAPTIVITY	Adaptation of Instruction
		TEACHINT	Perceived Teacher's Interest
	School-oriented variables	PERCOMP	Perception of Competitiveness at School (WLE)
		PERCOOP	Perception of Cooperation at School (WLE)
		ATTLNACT	Attitude Toward School: Learning Activities (WLE)
	School Climate	BELONG	Subjective Well-being: Sense of Belonging to School (WLE)
		BEINGBULLIED	Student's Experience of Being Bullied (WLE)

within each group. This method leverages the 'simputation' package in R, facilitating a tailored approach to missing value interpolation that accounts for inter-school variability. Such a strategy mitigates the inaccuracies that might arise from a more generalized data interpolation. Following this meticulous data preprocessing, the final sample encompasses a cohort of 12,058 students distributed across 361 schools.

Model introduction

Random forest model. The Classification and Regression Tree (CART) methodology, initially introduced by Breiman,[34] represents a seminal approach in decision tree algorithms. This technique employed a tree structure to model the intricate relationships between various features and response variables. It utilized recursive partitioning to segregate the dataset into distinct, non-overlapping subsets, each characterized by unique feature attributes. Notably, CART is versatile, supporting both classification and regression tasks in machine learning applications. In the context of this research, where the response variable is a reading achievement—a continuous measure—the CART regression tree model is particularly applicable, offering a tailored analytical framework for our investigation.

Breiman investigated the integrated learning algorithm of the Random Forest, which is built upon the Classification and Regression Tree (CART) model.[35] This algorithm includes both bagging and boosting methodologies. The Random Forest, a derivative of the bagging approach, employs stochastic sampling and variable selection techniques, which enhances the independence among the trees, allowing for a more diversified representation of the data and consequently reducing the generalization error's upper bound.[36] When compared to the traditional CART model, the Random Forest algorithm exhibits several advantages: (1) It mitigates the risk of overfitting; (2) It offers flexibility, maintaining accuracy even when data is partially missing; (3) It facilitates the determination of feature importance. However, there are notable drawbacks to this approach: (1) It can be time-consuming; (2) It requires substantial data storage resources; (3) The interpretability of its results is less straightforward than that of a single decision tree; (4) It may not adequately address the mixed effects in hierarchical data.

Integrated RE-EM regression tree model. The Random Effect-Expectation Maximization (RE-EM) regression tree represents an innovative synthesis of a Mixed-Effect linear model and a Classification and Regression Tree (CART) regression tree. This approach, pioneered by Sela and Simonoff,[32] integrates the conventional linear fixed-effects component of the mixed-effect model with a more dynamic regression tree framework. Crucially, it estimates the random-effects term using an Expectation Maximization (EM) algorithm. This integration is particularly effective in managing hierarchical data structures, offering significant explanatory power. The EM algorithm's incorporation into this model enhances its ability to accurately capture and analyze complex data relationships, a critical advantage in multilevel modeling scenarios.

The integration learning represents a significant domain in machine learning, predicated on the concept of synthesizing a more accurate classifier by amalgamating multiple individual classifiers into a cohesive, integrated system. This approach not only preserves the strengths of each original classifier but also minimizes the bias inherent in singular classifier predictions. Integration learning can be broadly categorized into three methodologies: Bagging, Boosting, and Stacking. Within this framework, Random Forest is classified under integration learning.

Table 3. Gender descriptive statistics

Student Gender	Number	Proportion
Male	6283	52.1
Female	5775	47.9
Total	12058	100

The fundamental principle of Random Forest involves the consolidation of numerous "weak" classifiers to enhance overall classifier efficacy. It specifically employs the Bagging technique, which involves the following steps: (1) Drawing *n* training samples with replacement from the original training set to constitute a new training set for each iteration; (2) Generating *M* sub-models by training on these newly formed training sets; and (3) In classification tasks, the final category is determined by a majority vote among the sub-models, while in regression tasks, a simple averaging method is employed to derive the predicted value.[37] At its core, Random Forest consists of a multitude of decision trees, each contributing to the collective predictive strength of the ensemble.

This study leverages the Bagging concept to construct an integrated Random Effect-Expectation Maximization (RE-EM) regression tree model. At its core, this model utilizes the RE-EM regression tree as the foundational unit, assembling an array of decision trees to enhance predictive accuracy and robustness. While a singular RE-EM regression tree may exhibit bias toward including all samples and features from the training data, potentially leading to overfitting, the incorporation of multiple, independent decision trees in a random forest framework significantly mitigates this risk. This is achieved through a reduction in variance and prediction error, thereby preventing overfitting.

The algorithmic construction of the RE-EM Random Forest model from the decision tree logic entails several key steps. Initially, the Random Forest parameters are defined, including the number of decision trees (typically set to 500 in a random forest), the number of samples, and the number of variables. The construction process then unfolds in stages: (1) creation of a *List* to store decision trees; (2) utilization of the *Sample* function to generate random subsets of samples with replacement; (3) application of the *Sample* function again to create subsets of random variables; (4) employing the *RE-EM tree* function to build a decision tree for each sampled set; and (5) storage of these decision trees in a list. The final stage involves prediction using the random forest, where the average outcome of all decision tree predictions is computed through the bagging method.

Interpretability of machine learning

Importance of features. The analysis of feature importance plays a crucial role in elucidating the relative significance of various variables in a model's predictive capability. This study utilizes DALEX, a prominent package in R, and is adept at providing a comprehensive explanation of the constructed model. Its capabilities extend beyond the mere assessment of model performance and also offer an in-depth analysis of the influence exerted by different feature variables on the response variable. This approach is instrumental in enhancing the understanding of the model's predictive dynamics and the specific contributions of each variable.

PDP plot. The Partial Dependence Plot (PDP) serves as a crucial tool for illustrating the interaction between explanatory variables and elucidating the underlying mechanisms of such interactions. A notable application of the PDP is in feature filtering: when the PDP curve of a feature is nearly horizontal or displays irregular fluctuations, it may indicate that the feature has minimal or no useful predictive power. Conversely, a steep PDP curve suggests a substantial contribution of the feature to the model, highlighting its relative importance. This dichotomy in PDP curve patterns offers a nuanced understanding of feature relevance within the predictive model.

RESULTS

Data preprocessing

During the PISA 2018 assessment, the original dataset exhibited instances of missing values. To address these gaps, this study adopts a method of grouping schools and then employs median interpolation within these groups for missing value imputation. This approach can mitigate the potential errors that might arise from applying a uniform method of imputation across the entire dataset since there are distinct differences between schools.

Descriptive statistics

The analysis of the data employs descriptive statistics to illustrate the trends in concentration and dispersion. Additionally, Analysis of Variance (ANOVA) explores the relationship between reading achievement scores and various categorical variables. A primary focus is on the categorical variable of gender, with the results presented in Table 3. The sample comprises 15-year-old students from four provinces/municipalities in China, including 6,283 male and 5,775 female students. This distribution results in a male-to-female ratio of 1.09, which is approximately equal to 1, indicating a relatively balanced gender ratio among the students from these regions participating in the assessment.

In the dataset released by PISA, the Weighted Likelihood Estimates (WLE) of various variables are standardized. An average greater than zero indicates a performance above the average level of OECD countries (regions), while a value less than zero signifies a performance below this average. This distinction is critical for interpreting the data. According to Table 4, the students from the four Chinese provinces/municipalities under study generally score slightly below the OECD average in several aspects, namely Socioeconomic and Cultural Status (ESCS),

Table 4. Descriptive statistics for continuous variables

Variables	Sample size	Mean	Standard deviation	Minimum	Maximum
DURECEC	12058	3.13	0.81	0.00	7.00
BSMJ	12058	68.33	15.84	11.01	88.96
MMINS	12058	282.8	131.75	0.00	2000.00
LMINS	12058	266.11	116.44	0.00	2400.00
SMINS	12058	323.2	213.99	0.00	2210.00
TMINS	12058	1896	415.09	150.00	3000.00
ESCS	12058	−0.36	1.09	−5.10	3.10
UNDREM	12058	0.20	0.99	−1.64	1.50
METASUM	12058	−0.12	0.96	−1.72	1.36
METASPAM	12058	0.09	0.96	−1.41	1.33
DISCLIMA	12058	0.81	1.03	−2.71	2.03
TEACHSUP	12058	0.42	0.88	−2.71	1.34
DIRINS	12058	0.51	1.02	−2.94	1.82
PERFEED	12058	0.35	1.03	−1.60	2.00
EMOSUPS	12058	0.01	0.93	−2.45	1.03
STIMREAD	12058	0.64	1.03	−2.30	2.09
ADAPTIVITY	12058	0.43	1.04	−2.27	2.01
TEACHINT	12058	0.38	0.97	−2.22	1.82
JOYREAD	12058	0.98	0.84	−2.70	2.70
SCREADCOMP	12058	0.08	0.86	−2.44	1.88
SCREADDIFF	12058	0.12	0.95	−1.89	2.78
PERCOMP	12058	0.16	0.95	−1.99	2.04
PERCOOP	12058	0.23	1.00	−2.14	1.68
ATTLNACT	12058	0.16	0.92	−2.54	1.08
COMPETE	12058	0.42	0.82	−2.35	2.01
WORKMAST	12058	0.30	0.89	−2.74	1.82
GFOFAIL	12058	0.01	0.87	−1.90	1.90
EUDMO	12058	0.09	0.92	−2.15	1.74
SWBP	12058	0.10	0.89	−3.07	1.24
RESILIENCE	12058	−0.07	0.95	−3.17	2.37
MASTGOAL	12058	0.06	0.91	−2.50	1.90
BELONG	12058	−0.15	0.91	−3.26	2.76
BEINGBULLIED	12058	−0.24	0.88	−0.78	3.86
PV1READ	12058	561.3	90.34	208.22	847.85

Meta-cognition: Summarizing (METASUM), Resilience (RESILIENCE), Subjective Well-being: Sense of Belonging to School (BELONG), and Student's Experience of Being Bullied (BEINGBULLIED). However, it is notable that the average value of Joy/Like Reading (JOYREAD) for these students is 0.98, surpassing the OECD average. Furthermore, in the context of Chinese education, certain classroom dynamics such as Disciplinary Climate in Test Language Lessons (DISCLIMA), Teacher's Stimulation of Reading Engagement Perceived by Student (STIMREAD), and Teacher-Directed Instruction (DIRINS) demonstrate mean values of 0.81, 0.63, and 0.51, respectively. These figures also exceed the average observed in OECD countries (regions), indicating distinctive features of the educational environment in these Chinese provinces/municipalities.

Through one-way ANOVA to investigate differences in reading achievement across school categoriesr, with findings presented in Table 5. The results for schools reveal a *p*-value approximately equaling to 0, indicating statistically significant disparities in reading achievement among students across the schools in China, which supports our idea that the normal machine learning model is not suitable for our data and the RE-EM model is necessary.

The preprocessed dataset, comprising 12,058 samples with 35 feature variables and one output variable, is partitioned randomly into two subsets. This division follows a 70:30 ratio, where 70% of the data forms the training set and the remaining 30% constitutes the test set. The principal outcomes of each model are summarized as follows.

Table 5. Analysis of variance of students' reading achievement on schools in four provinces/municipalities in China

Source of Variation	SS	df	MS	F	P
Between groups	47302955.19	360	131397.10	30.08	0.000***
Within groups	51095325.75	11697	4368.24		
Total	98398280.94	12057			

Note: "***," "**," and "*" indicate significance at the 0.1%, 1%, and 5% levels, respectively.

Model results

Model 1: Classification and regression tree model

Figure 1 presents the estimated regression tree, elucidating the outcomes of the regression tree analysis, which includes a box-and-whisker plot beneath each branch of the tree. The most pivotal variable segmented by the regression tree model is Meta-cognition: Assess Credibility (METASPAM), with a division threshold set at −0.04. Examination of the box-and-whisker plots reveals that the data subsets, corresponding to each segmentation criterion, display an approximately symmetric distribution.

In terms of reading achievement scores, it can be concluded that Meta-cognition: Assess Credibility (METASPAM) is the most influential variable, with subsequent significant variables being Science Learning Time (SMINS), Socioeconomic and Cultural Status (ESCS), Joy/Like Reading (JOYREAD), and Work Mastery (WORKMAST), in descending order of impact (the other variables examined do not exert a significant influence on reading achievement scores). The PISA 2018 dataset categorizes lesson durations into four primary types: Total Learning Time (TMINS), Math Learning Time (MMINS), Reading Learning Time (LMINS), and Science learning time (SMINS). According to the results derived from the Classification and Regression Tree (CART) model, all categories, except for MMINs, significantly influence reading achievement scores, which underscores the pivotal role of time spent in educational activities at school in enhancing students' reading proficiency.

Model 2: Random forest model

For the variables needed to calculate the importance of variables in a Random Forest, the original data is replaced with randomly generated data, and the model accuracy or GINI coefficient reduction index is further calculated. The model accuracy, particularly for out-of-bag (OOB) samples, is determined by the average decline in accuracy, serving as a method for assessing feature importance. Meanwhile, the GINI coefficient method offers insight into the relative significance of each feature, where the average across all trees in the random forest quantifies the importance of individual features.

Figure 2 illustrates the feature importance ranking within the random forest model, offering insights from two distinct analytical perspectives. When evaluated in terms of model accuracy, Meta-cognition: Assess Credibility (METASPAM) emerges as the most important variable, followed in importance by Socioeconomic and Cultural Status (ESCS), Science Learning Time (SMINS), Meta-cognition: Summarizing (METASUM), Joy/Like Reading (JOYREAD), Eudaemonia: Meaning in Life (EUDMO), Reading Learning Time (LMINS), Total Learning Time (TMINS), Math Learning Time (MMINS), Student's Expected occupational status (BSMJ), and Self-Concept of Reading: Perception of Competence (SCREADCOMP) (variables beyond these exhibit negligible effects on reading achievement scores). From the GINI coefficient viewpoint, METASPAM retains the highest significance, succeeded by ESCS, METASUM, SMINS, JOYREAD, BSMJ, and TMINS, with other variables demonstrating minimal impact on reading achievement scores.

Model 3: Random effect-expectation maximization regression tree model

The random effect-expectation maximization (RE-EM) regression tree model is operationalized through the use of the "REEMtree" package in R, which is a sophisticated data mining methodology specifically designed for analyzing longitudinal and clustered data. It integrates the hierarchical architecture of a mixed-effect model with the precision of a tree-based estimation method. Figure 3 displays the estimated regression tree derived from the application of the "REEMtree" function.

The results derived from the RE-EM regression tree analysis indicate that Meta-cognition: Assess Credibility (METASPAM) emerges as the most significant characteristic. The following, in descending order of importance, by Meta-cognition: Summarizing (METASUM), Joy/Like Reading (JOYREAD), Science Learning Time (SMINS), and Socioeconomic and Cultural Status (ESCS). (The impact of the remaining variables on the model is comparatively insignificant.).

Model 4: Integrated random effect-expectation maximization regression tree model

In advancing the random effect-expectation maximization (RE-EM) regression tree methodology, this study integrates it into a random forest framework, utilizing the logic of "through decision trees to construct random forest." Consequently, an optimized random forest (RF) model, grounded in the principles of the RE-EM regression tree, is constructed. Analysis based on this refined model reveals a hierarchy of feature importance. At the pinnacle of this hierarchy is the Meta-cognition: Assess Credibility (METASPAM) feature, demonstrating the highest significance, and sequentially followed by Meta-cognition: Summarizing (METASUM), Joy/Like Reading (JOYREAD), Science Learning Time (SMINS), Socioeconomic and Cultural Status (ESCS), Meta-cognition: understanding and remembering (UNDREM), Eudaemonia: Meaning in Life (EUDMO), Self-Concept of Reading: Perception of Difficulty (SCREADDIFF), Self-Concept of Reading: Perception of Competence

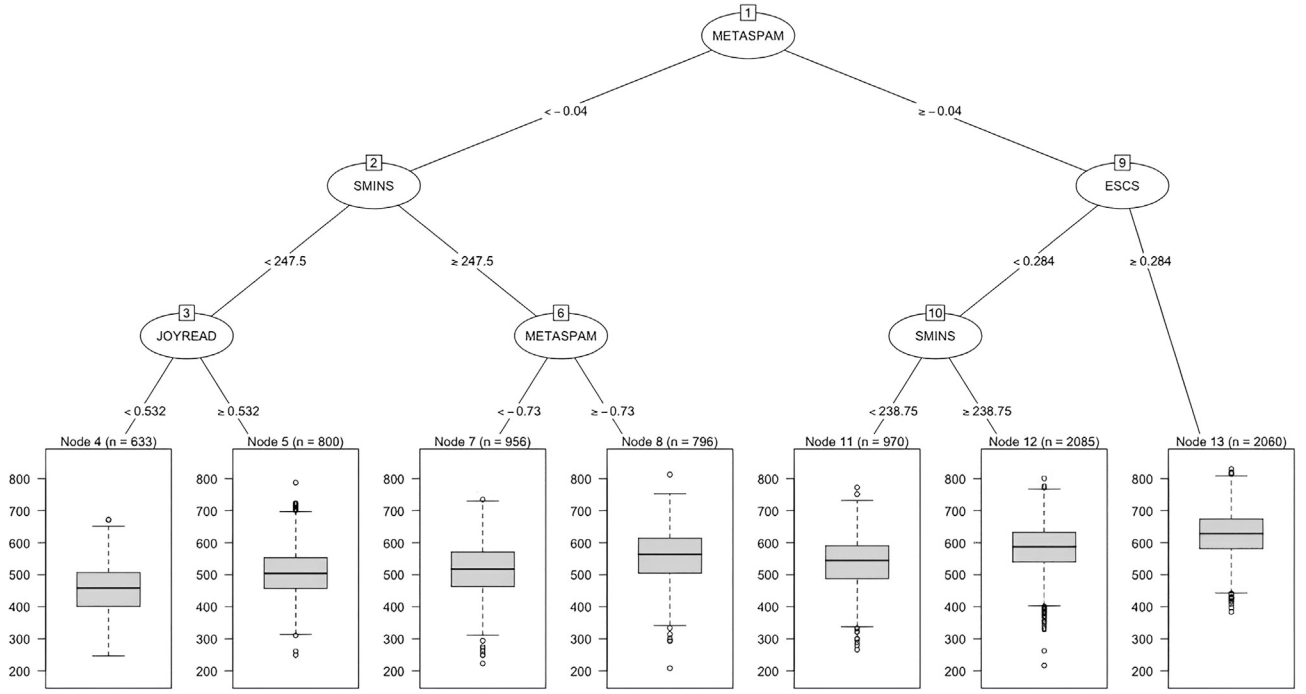


Figure 1. CART regression tree model

(SCREADCOMP), and Total Learning Time (TMINS). (Features beyond these are found to exert a negligible impact on the outcomes of the model).

Model comparison

To assess the potential for overfitting in the model, a test set is employed. This involves an iterative process of training and testing the model to evaluate its accuracy. The accuracy is primarily judged based on the root-mean-square error (RMSE), which serves as a key metric for quantifying the deviation between observed (true) values and predicted values and effectively reflects the degree of dispersion in these values. The formula of RMSE is presented later in discussion:

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(observed_t - predicted_t\right)^2} \quad \text{(Equation 1)}$$

Where the $observed_t$ refers to the observed values, $predicted_t$ refers to the predicted values. The results of the RMSE of each model are shown in Table 6, in which the integrated RE-EM regression tree model has the smallest RMSE, indicating the best prediction performance.

Exploration of the mechanism of influence of important factors

This study establishes that the integrated RE-EM regression tree model exhibits superior predictive performance in assessing reading achievement. Utilizing this model, the study identifies Meta-cognition: Assess Credibility (METASPAM) as the most influential variable, followed by Meta-cognition: Summarizing (METASUM), and Joy/Like Reading (JOYREAD), respectively. Subsequent analysis, employing Partial Dependence Plots (PDP), will delve into the specific mechanisms through which METASPAM, METASUM, and JOYREAD impact reading achievement.

(1) The Relationship between Reading Achievement and Meta-cognition: Assess Credibility (METASPAM)

The analysis of the PDP of reading achievement scores in relation to Meta-cognition: Assess Credibility (METASPAM) is presented in Figure 4. The plot features a zigzagging line pattern, indicative of a nonlinear relationship between reading scores and METASPAM. Notably, reading scores exhibit increased variability when METASPAM values fall below 0, while fluctuations tend to level off as METASPAM values rise above 0. The overall trajectory of the lines shifts toward the upper right, suggesting that higher METASPAM positively correlates with improved reading achievement scores. However, the marginal positive effect appears to diminish as METASPAM values increase, and results in a ceiling effect. From a practical standpoint, this finding implies that to enhance reading achievement through the development of METASPAM, the most effective strategy is to cultivate students' METASPAM to align with the average level observed across the student

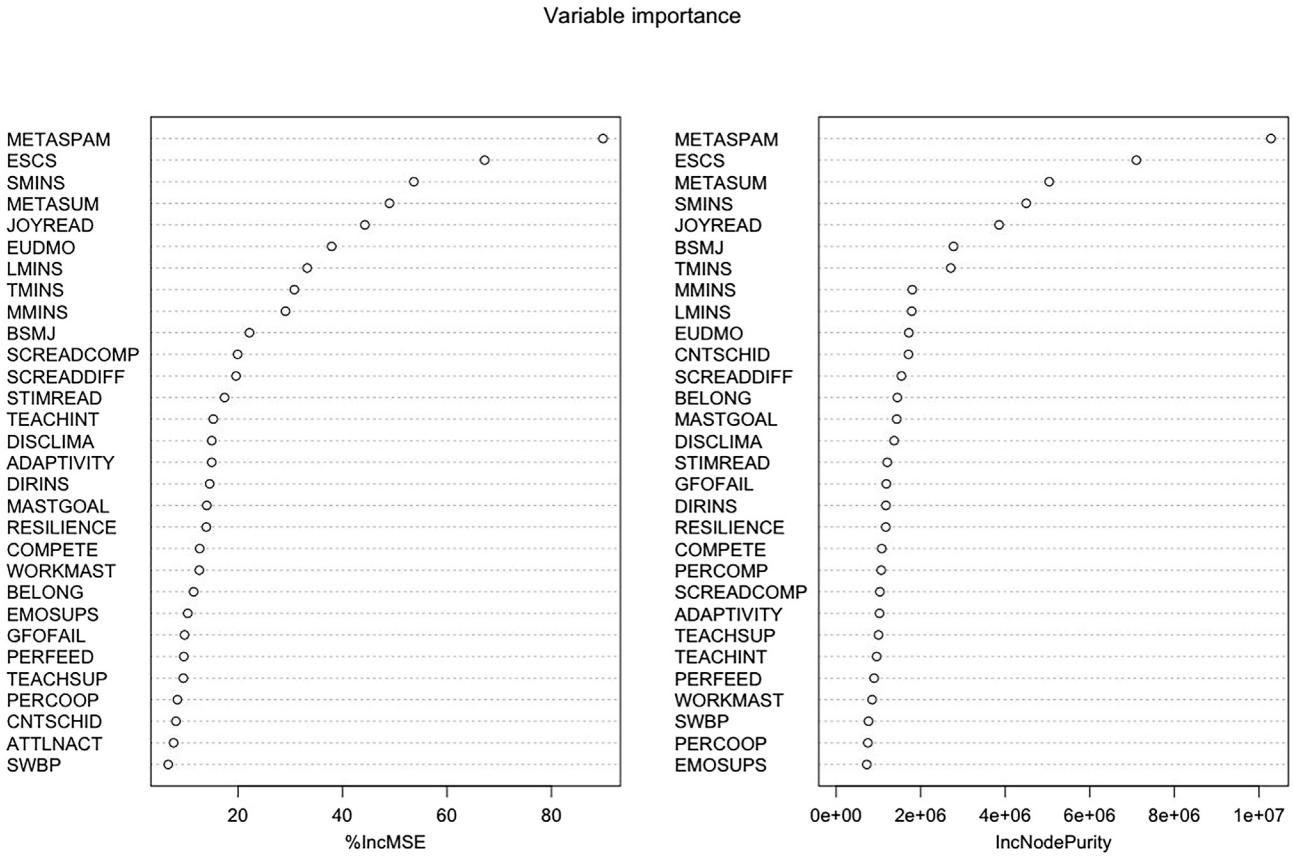


Figure 2. Random forest feature importance ranking plot

population. Advancing METASPAM skills beyond this average is likely to yield diminishing returns, or potentially even a decline, in reading achievement improvements. Therefore, educators and educational institutions should focus on reducing disparities in METASPAM among students, particularly by supporting those whose METASPAM is currently below the average.

(2) The Relationship between Reading Achievement and Meta-cognition: Summarizing (METASUM)

The PDP of reading achievement in relation to Meta-cognition: Summarizing (METASUM), as depicted in Figure 5. The analysis reveals a significant nonlinear correlation between reading scores and METASUM. Specifically, METASUM exhibits fluctuations around a score of 540 when its value is below −1. In the range of −0.5 to 0, reading achievement scores demonstrate considerable volatility, characterized by erratic increases and decreases, suggesting instability in this interval. Conversely, for METASUM values exceeding 0, reading scores tend to stabilize,

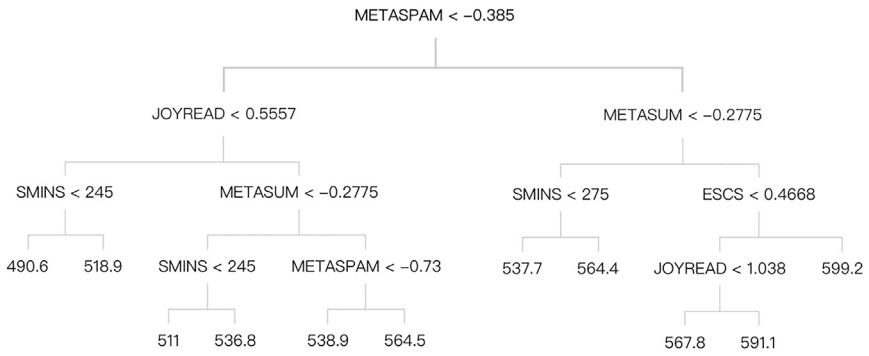


Figure 3. RE-EM regression tree model

Table 6. Root means square error results (RMSE) in each model

Model	RMSE
CART Regression Tree Model	75.33
RF Model	62.89
RE-EM Regression Tree Model	61.64
Integrated RE-EM Regression Tree Model	60.11

showing minor fluctuations around an average of 570. The overall trend suggests a positive impact of METASUM on reading achievement scores. However, this positive effect appears to attenuate when METASUM surpasses the mean value. Overall, the lower or higher values of METASUM imply smaller marginal benefits, whereas the marginal benefits are more pronounced within the range of −1 to 0, demonstrating an S-shaped trend. Notably, METASUM levels near the mean are associated with a broad spectrum of reading achievement scores. This finding has practical implications: optimizing students' METASUM to improve reading achievement is most effective when focusing on developing METASUM to approximate the average level across the student population. Advancing METASUM skills beyond this average may lead to reduced, or even negative, gains in reading achievement. Consequently, educators and institutions should aim to narrow the METASUM gap among students, with particular emphasis on enhancing the skills of those whose METASUM is below average.

(3) Investigating the relationship between of reading achievement and Joy/Like Reading (JOYREAD)

Figure 6 presents a PDP illustrating the relationship between reading achievement and Joy/like Reading (JOYREAD). The plot reveals significant fluctuations in the lines, indicating a pronounced nonlinear correlation between reading scores and JOYREAD. Instances where JOYREAD is less than 0, denoting lower levels of reading enjoyment, correspond to an ambiguous influence on reading scores, with a diminished explanatory power for reading performance. Conversely, in scenarios where JOYREAD exceeds 0, the general trajectory of the lines suggests a positive impact of reading enjoyment on reading scores. When the JOYREAD exceeds 1, reading scores remain at a high level, indicating a ceiling effect. Consequently, the overall relationship between JOYREAD and reading achievement exhibits an S-shaped curve. The marginal benefits of increasing JOYREAD on reading scores are highest when it is between 0 and 1.

DISCUSSION AND CONCLUSION

This study, utilizing data from four provinces/municipalities in China as part of PISA 2018, develops and compares several models: the CART regression tree model, the random forest model, the RE-EM regression tree model, and an integrated RE-EM regression tree model. The aim is to contrast the predictive performance of traditional machine learning approaches with that of the random effect-expectation maximization (RE-EM) method. The findings indicate that the integrated RE-EM regression tree model exhibits superior predictive performance. Moreover, this study delves into the critical factors influencing reading achievement and their mechanisms through the integrated RE-EM regression tree model. Based on the three-dimensional framework we constructed, this study compares the impact of variables at the individual, school, and family levels on reading achievement, confirming the pivotal role of individual-level variables in students' reading achievement. The study identifies that three types of Meta-cognition and Joy/Like Reading are the most significant factors influencing reading achievement, which is consistent with the findings[38,39] of Hao et al. and Bu & Chen, both of which emphasize the critical role of Meta-cognition strategies and Joy/Like Reading in reading achievement. These variables, which belong to the individual cognitive level, exhibit a clear nonlinear relationship with reading achievement, highlighting the advantages of the RE-EM method. This study reveals the previously unrecognized nonlinear impacts of Meta-cognition strategies and Joy/Like Reading on reading achievement. The importance of the most critical school-level variables, study time, and the most significant family-level variable, socioeconomic and cultural status, is found to be lower than that of Meta-cognition. The following is a more detailed discussion of the aforementioned conclusions.

(1) Among the four models evaluated, the integrated RE-EM regression tree model demonstrates the most effective predictive performance, particularly when considering the variation in student performance across different school levels. It can be discerned that when employing hierarchically structured data, such as PISA, for data mining to explore the interactions among variables, it is essential to conduct grouped studies of schools. The disparities between schools, considered as fixed effects, play a crucial role in influencing reading achievement.

(2) Employing the integrated RE-EM regression tree model, this study ranks the following factors in order of importance for feature selection: Meta-cognition: Assess Credibility (METASPAM), Meta-cognition: Summarizing (METASUM), Joy/Like Reading (JOYREAD), Science Learning Time (SMINS), Socioeconomic and Cultural Status (ESCS), Meta-cognition: Understanding and Remembering (UNDREM), and Eudaemonia: Meaning in Life (EUDMO). METASPAM and METASUM are consistently identified across all models as key factors influencing students' reading achievement, with a positive effect on scores. Meta-cognition strategies refer to an individual's awareness of effective strategies, requiring the ability to recognize and understand when various strategies are most effective, and to select and adjust these strategies based on personal conditions to achieve goals.[40] Numerous studies have demonstrated that Meta-cognition strategies can effectively predict reading achievement.[41–43] The CART and random forest models emphasize the influence of ESCS and SMINS, while the integrated and regular RE-EM regression tree models highlight the impact of JOYREAD. The

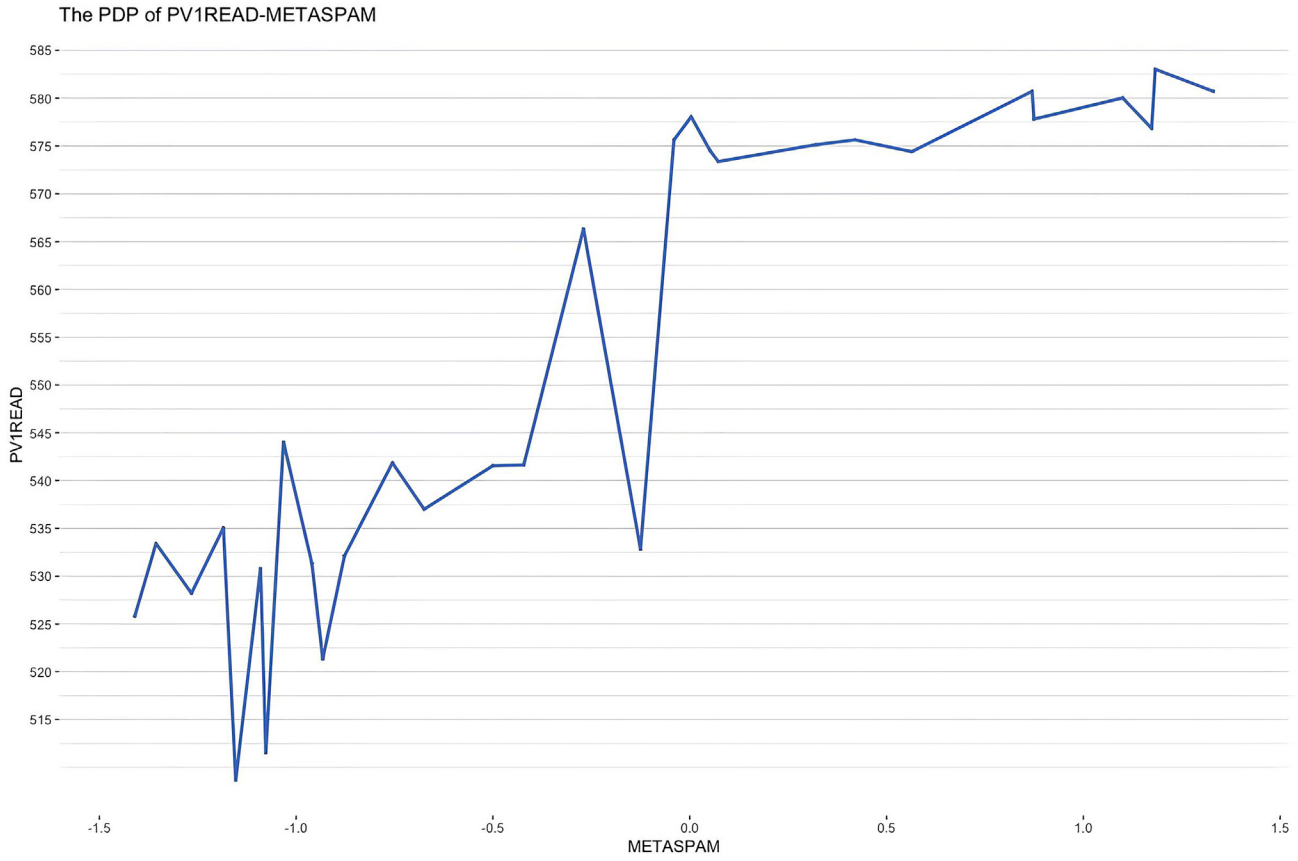


Figure 4. Partial dependence plot between meta-cognition: assess credibility and reading achievement

random forest model uniquely considers EUDMO as an important variable, whereas the other models do not attribute special significance to it in affecting reading achievement.

(3) In analyzing the effects of Meta-cognition: Assess Credibility (METASPAM), Meta-cognition: Summarizing (METASUM), and Joy/Like Reading (JOYREAD), all models agree that METASPAM exerts the strongest influence on reading achievement, exhibiting a non-linear relationship, which indicates the limitations of traditional linear regression methods in capturing non-linear impacts of explanatory variables. However, traditional studies predominantly utilized linear models.[44,45] Furthermore, the impact of METASUM on reading achievement is variable, showing larger score differences for minor variations in METASUM. The effects of reading Meta-cognition and Joy/Like Reading on reading achievement exhibit nonlinear relationships. Specifically, Meta-cognition: Summarizing and Joy/Like Reading show S-shaped curves in their influence on reading achievement, while Meta-cognition: Assessing Credibility demonstrates diminishing marginal returns. Overall, these variables display a clear ceiling effect on reading achievement, where the marginal effects diminish once the variable exceeds a certain threshold. This phenomenon can be attributed to the multifaceted nature of reading achievement, which is influenced by various interacting factors rather than solely by Meta-cognition and Joy/Like Reading. For instance, students' Meta-cognition strategies can impact reading achievement through cognitive strategies[46] and by enhancing student engagement.[47] Similarly, Joy/Like Reading influences reading achievement through attention allocation[48] and reading engagement.[49] When students exhibit high levels of Joy/Like Reading and Meta-cognition, further improvements in reading achievement necessitate greater contributions from other mediating variables.

This study underscores the significance of students' reading strategies and emotional engagement in their reading achievement. Consequently, educators should acknowledge that teachers must not only promote reading among students but also instruct them on effective reading techniques, which involves equipping students with advanced reading strategies and meta-cognition skills. Beyond cultivating students' interest in reading though pedagogical practices, it is essential to provide comprehensive support encompassing emotional care, instructional guidance, and skill development. When selecting reading materials, it is crucial to consider their engaging nature and align them with students' interests. By adopting an interest-driven approach, educators can facilitate students' understanding of the relevance of the content, thereby elevating their reading proficiency. Guiding students to master effective reading strategies and cultivating their positive reading attitude toward reading are the most effective ways to improve students' reading achievement.

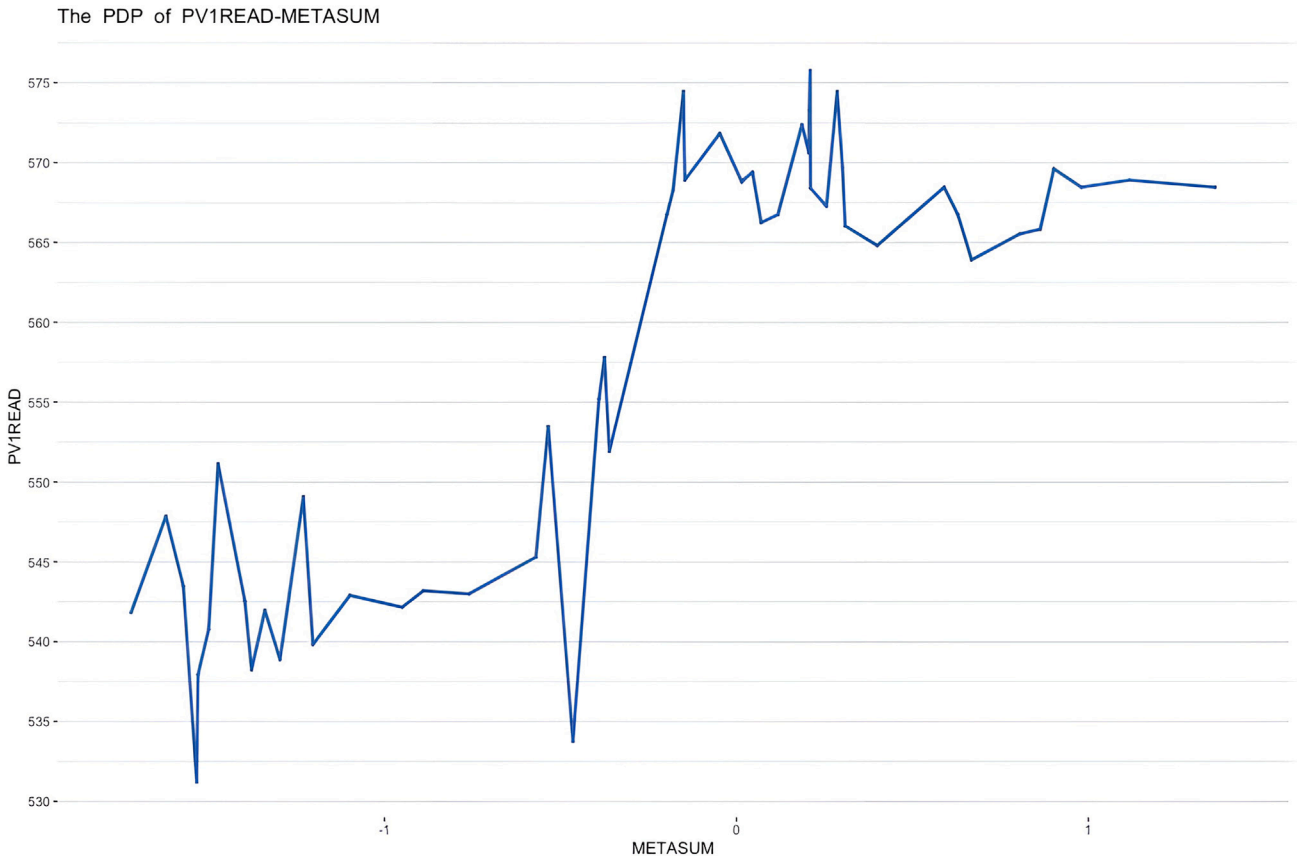


Figure 5. Partial dependence plot between meta-cognition: summarizing and reading achievement

Weaknesses and prospects

This study introduces the integrated Random Effect-Expectation Maximization (RE-EM) regression tree model, developed following the foundational principles of the random forest model. The model comprises an ensemble of multiple decision trees, designed to handle continuous response variables. However, its applicability to discrete response variables necessitates further investigation. Moreover, this research establishes a random forest model based on the RE-EM regression tree. While the RE-EM tree serves as the fundamental decision tree in this context, its potential integration with other algorithmic approaches warrants additional comparative analyses. Future research endeavors should aim to construct a more comprehensive and rigorous random forest model based on the RE-EM regression tree. Such development endeavors to align with traditional machine learning paradigms, thereby paving the way for more in-depth and systematic investigations.

The RE-EM regression tree model presents an opportunity for extension to more comprehensive integrated algorithms, enhancing the efficacy of machine learning models. Future research avenues may include exploring novel applications of data mining techniques to nested data. This exploration could entail a more synergistic integration of linear regression with machine learning methodologies, leveraging the strengths of both approaches. Such a convergence aims to augment the theoretical foundations of data mining methods. The objective is to refine existing machine learning models by incorporating more sophisticated linear regression techniques, thereby enriching their analytical capabilities and precision.

In the end, constrained by the PISA data, this study does not include a sufficient number of family-level (aside from socioeconomic and cultural status) and societal-level variables. Family-level variables, such as parental involvement, parental accompaniment, and parenting styles, have been shown to significantly impact students' reading achievement. However, PISA does not collect parental questionnaire information in China, resulting in the absence of these variables in the Chinese dataset. Furthermore, more macro-level variables, such as home-school interaction, societal values, and local economic development levels, also influence students' reading abilities. Future research will include these variables to explore their effects on reading achievement.

RESOURCE AVAILABILITY

Lead contact

For additional details and inquiries regarding data sharing, please contact HaoLiu (liuhao@bnu.edu.cn), who will handle and fulfill your requests promptly.

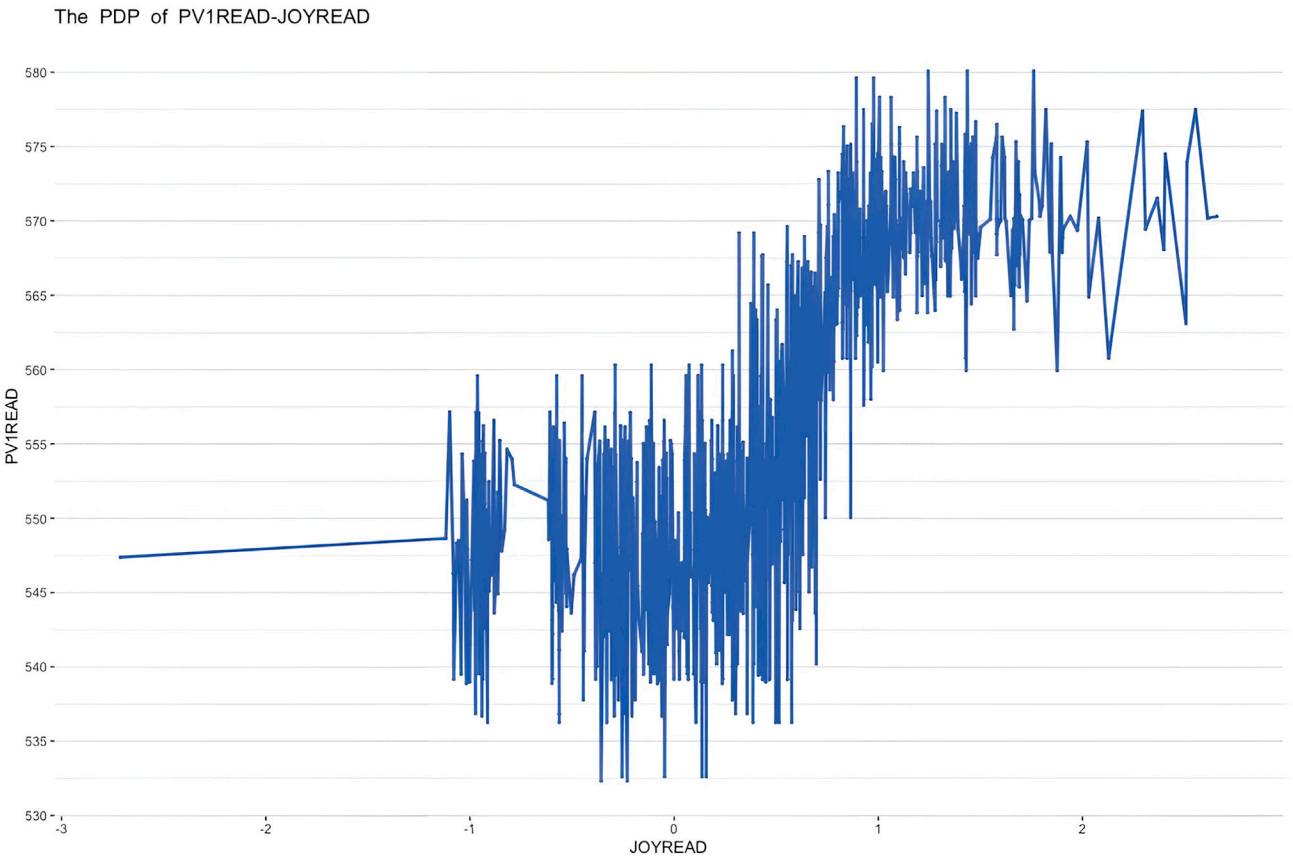


Figure 6. Partial dependence plot between joy/like reading and reading achievement

Materials availability

This study did not generate any new unique materials other than the data collected.

Data and code availability

- The datasets are available in the following repository and publicly accessible: https://www.oecd.org/en/data/datasets/pisa-2018-database.
- The codes used to generate model are available in the following repository and publicly accessible: https://cran.r-project.org/web/packages/REEMtree/index.
- Any additional information required to reanalyze the data reported in this work article is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported in part by the National Social Science Fund of China under Grant 20CTJ019.

AUTHOR CONTRIBUTIONS

Conceptualization, HaoLiu and Dongxia Yang; methodology, Dongxia Yang and XiChen; software, Dongxia Yang and XiChen; formal analysis, Dongxia Yang; writing - original draft, Dongxia Yang; writing - review and editing, HaoLiu, Shangran Nie; writing - polishing and revising the article, Shangran Nie; supervision, HaoLiu; project administration, HaoLiu; funding acquisition, HaoLiu.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

Received: January 31, 2024
Revised: June 5, 2024
Accepted: August 27, 2024
Published: August 31, 2024

REFERENCES

1. OECD (2019). PISA 2018 Assessment and Analytical Framework (OECD Publishing). https://doi.org/10.1787/b25efab8-en.
2. OECD (2010). PISA 2009 Assessment Framework (OECD Publishing). https://doi.org/10.1787/9789264062658-en.
3. Li, W.N. (2019). Simulation Study and Evaluation of Random Effect-Expectation Maximization Regression Tree Model (Guangdong Pharmaceutical University). https://doi.org/10.27690/d.cnki.ggdyk.2019.000015.
4. Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. Am. Psychol. *32*, 513–531. https://doi.org/10.1037/0003-066x.32.7.513.
5. Hu, J., and Wang, M.S. (2022). Studies on Influencing Factors Related to Students' Reading Literacy Based on HLM—Evidence from PISA 2018 in China. Inf. Sci. *40*, 127–132+140. https://doi.org/10.13833/j.issn.1007-7634.2022.02.017.
6. Liao, Q., and Wang, Z. (2020). Chinese Students' Reading Literacy and the Related Factors—An Analysis Based on PISA 2018 Data. J. Shanghai Edu. Res. *06*, 24–29. https://doi.org/10.16194/j.cnki.31-1059/g4.2020.06.006.
7. Logan, S., Medford, E., and Hughes, N. (2011). The importance of intrinsic motivation for high and low ability readers' reading comprehension performance. Learn. Indiv Differ *21*, 124–128. https://doi.org/10.1016/j.lindif.2010.09.011.
8. Chen, C.J. (2020). New Findings in Reading Literacy Assessment among Students in the Four Provinces/Municipalities of China in PISA 2018. J. East China Normal Univ. (Edu. Sci.) *38*, 22–62. https://doi.org/10.16382/j.cnki.1000-5560.2020.05.002.
9. Schoor, C. (2016). Utility of reading — Predictor of reading achievement? Learn. Indiv Differ *45*, 151–158. https://doi.org/10.1016/j.lindif.2015.11.024.
10. Nalipay, M.J.N., King, R.B., and Cai, Y. (2020). Autonomy is equally important across East and West: Testing the cross-cultural universality of self-determination theory. J. Adolesc. *78*, 67–72. https://doi.org/10.1016/j.adolescence.2019.12.009.
11. Liu, B.C., and Kang, Y.F. (2021). The Influence of School Background on Students' Achievement: A Comparative Analysis between China(B-S-J-Z) and Countries with High Scores in PISA 2018. Glob. Edu. *50*, 44–62.
12. Berkowitz, R. (2021). School climate and the socioeconomic literacy achievement gap: Multilevel analysis of compensation, mediation, and moderation models. Child. Youth Serv. Rev. *130*, 1–10. https://doi.org/10.1016/j.childyouth.2021.106238.
13. Jia, Y., and Zhang, J.H. (2020). Interpretation of PISA 2018: Analysis of the Current Situation of Teacher' Classroom Teaching in Four Provinces/Municipalities of China—Analysis and International Comparison Based on Four Provinces/Municipalities of China PISA 2018 Data. School Adm. *1*, 16–20.
14. Sadoughi, M., and Hejazi, S.Y. (2021). Teacher support and academic engagement among EFL learners: The role of positive academic emotions. Stud. Educ. Eval. *70*, 101060. https://doi.org/10.1016/j.stueduc.2021.101060.
15. Law, Y.K. (2011). The role of teachers' cognitive support in motivating young Hong Kong Chinese children to read and enhancing reading comprehension. Teach. Teach. Educ. *27*, 73–84. https://doi.org/10.1016/j.tate.2010.07.004.
16. Ning, B., Van Damme, J., Van Den Noortgate, W., Yang, X., and Gielen, S. (2015). The influence of classroom disciplinary climate of schools on reading achievement: a cross-country comparative study. Sch. Effect. Sch. Improv. *26*, 586–611. https://doi.org/10.1080/09243453.2015.1025796.
17. Konstantopoulos, S., and Borman, G.D. (2011). Family Background and School Effects on Student Achievement: A Multilevel Analysis of the Coleman Data. Teach. Coll. Rec. Voice Scholarship Edu. *113*, 97–132. https://doi.org/10.1177/016146811111300101.
18. Banerjee, P.A. (2016). A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools. Cogent Edu. *3*, 1178441. https://doi.org/10.1080/2331186x.2016.1178441.
19. Camp, D. (2007). Who's Reading and Why: Reading Habits of 1st Grade Through Graduate Students. Read. Horiz. *47*, 251–268.
20. Celik, B. (2019). A Study on the Factors Affecting Reading and Reading Habits of Preschool Children. Int. J. Engl. Linguist. *10*, 101. https://doi.org/10.5539/ijel.v10n1p101.
21. Greaney, V., and Hegarty, M. (1987). Correlates of leisure-time reading. J. Res. Read. *10*, 3–20. https://doi.org/10.1111/j.1467-9817.1987.tb00278.x.
22. Chiu, M.M. (2018). Qatar family, school, and child effects on reading. Intern. J. Comparat. Edu. Develop. *20*, 113–127. https://doi.org/10.1108/ijced-03-2018-0004.
23. Netten, A., Voeten, M., Droop, M., and Verhoeven, L. (2014). Sociocultural and educational factors for reading literacy decline in the Netherlands in the past decade. Learn. Indiv Differ *32*, 9–18. https://doi.org/10.1016/j.lindif.2014.02.002.
24. Baker, L., Scher, D., and Mackler, K. (1997). Home and family influences on motivations for reading. Educ. Psychol. *32*, 69–82. https://doi.org/10.1207/s15326985ep3202_2.
25. Hofslundsengen, H., Gustafsson, J.-E., and Hagtvet, B.E. (2018). Contributions of the Home Literacy Environment and Underlying Language Skills to Preschool Invented Writing. Scand. J. Educ. Res. *63*, 653–669. https://doi.org/10.1080/00313831.2017.1420686.
26. Goldfeld, S., Moreno-Betancur, M., Guo, S., Mensah, F., O'Connor, E., Gray, S., Chong, S., Woolfenden, S., Williams, K., Kvalsvig, A., et al. (2021). Inequities in Children's Reading Skills: The Role of Home Reading and Preschool Attendance. Acad. Pediatr. *21*, 1046–1054. https://doi.org/10.1016/j.acap.2021.04.019.
27. Guryan, J., Hurst, E., and Kearney, M. (2008). Parental Education and Parental Time with Children. J. Econ. Perspect. *22*, 23–46. https://doi.org/10.1257/jep.22.3.23.
28. Xia, X. (2022). Family Income, Parental Education and Chinese Preschoolers' Cognitive School Readiness: Authoritative Parenting and Parental Involvement as Chain Mediators. Front. Psychol. *13*, 745093. https://doi.org/10.3389/fpsyg.2022.745093.
29. Huang, T., and Benoliel, P. (2023). Principal time use and student academic achievement in Singapore. Int. J. Educ. Manag. *37*, 1401–1424. https://doi.org/10.1108/IJEM-08-2023-0427.
30. Wu, X., and Zhang, Y. (2024). Effects of individual attributes, family background, and school context on students' global competence: Insights from the OECD PISA 2018. Int. J. Educ. Dev. *106*, 102996. https://doi.org/10.1016/j.ijedudev.2024.102996.
31. Lin, S., and Luo, W. (2019). A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. Multivariate Behav. Res. *54*, 578–592. https://doi.org/10.1080/00273171.2018.1552555.
32. Sela, R.J., and Simonoff, J.S. (2011). RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. *86*, 169–207. https://doi.org/10.1007/s10994-011-5258-3.
33. Fu, W., and Simonoff, J.S. (2015). Unbiased regression trees for longitudinal and clustered data. Comput. Stat. Data Anal. *88*, 53–74. https://doi.org/10.1016/j.csda.2015.02.004.
34. Breiman, L. (1984). Classification and Regression Trees (Wadsworth International Group).
35. Breiman, L. (2001). Random Forests. Mach. Learn. *45*, 1.
36. Cao, T.Y. (2018). Classification of Impurity Functions in Different Data Types Based on CART Model. Stat. Decision *34*, 77–79. https://doi.org/10.13546/j.cnki.tjyjc.2018.10.018.
37. Yuan, H.Y. (2022). Method of Multi-source Information Fusion of Indoor Positioning Based on Bluetooth and Particle Filter. Inform. Comput. *34*, 44–46.
38. Hao, L., Xi, C., and Xiaoxiao, L. (2022). Factors influencing secondary school students' reading literacy: An analysis based on XGBoost and SHAP methods. Front. Psychol. *13*, 1–18. https://doi.org/10.3389/fpsyg.2022.948612.
39. Bu, Y., and Chen, F. (2023). What key contextual factors contribute to students' reading literacy among top-performing countries and economies? Statistical and machine learning analyses. Int. J. Educ. Res. *122*, 102267. https://doi.org/10.1016/j.ijer.2023.102267.
40. Tomporowski, P.D., McCullick, B., Pendleton, D.M., and Pesce, C. (2015). Exercise and children's cognition: The role of exercise

characteristics and a place for metacognition. J. Sport Health Sci. *4*, 47–55. https://doi.org/10.1016/j.jshs.2014.09.003.

41. Muhid, A., Amalia, E.R., Hilaliyah, H., Budiana, N., and Wajdi, M.B.N. (2020). The Effect of Metacognitive Strategies Implementation on Students' Reading Comprehension Achievement. Int. J. InStruct. *13*, 847–862. https://doi.org/10.29333/iji.2020.13257a.
42. Qi, X. (2020). Effects of Self-Regulated Learning on Student's Reading Literacy: Evidence From Shanghai. Front. Psychol. *11*, 555849. https://doi.org/10.3389/fpsyg.2020.555849.
43. Lau, K., and Chan, D.W. (2003). Reading strategy use and motivation among Chinese good and poor readers in Hong Kong. J. Res. Read. *26*, 177–190. https://doi.org/10.1111/1467-9817.00195.
44. Ma, L., Luo, H., and Xiao, L. (2021). Perceived teacher support, self-concept, enjoyment and achievement in reading: A multilevel mediation model based on PISA 2018. Learn. Indiv Differ *85*, 101947. https://doi.org/10.1016/j.lindif.2020.101947.
45. Gu, Y.x., and Lau, K.l. (2023). Reading instruction and reading engagement and their relationship with Chinese students' PISA reading performance: Evidence from B-S-J-Z, Hong Kong, and Chinese Taipei. Int. J. Educ. Res. *120*, 102202. https://doi.org/10.1016/j.ijer.2023.102202.
46. Ghafournia, N., and Afghari, A. (2013). The Interaction between Reading Comprehension Cognitive Test-Taking Strategies, Test Performance, and Cognitive Language Learning Strategies. Procedia Soc. Behav. Sci. *70*, 80–84. https://doi.org/10.1016/j.sbspro.2013.01.041.
47. Lee, J., and Shute, V.J. (2010). Personal and Social-Contextual Factors in K–12 Academic Performance: An Integrative Perspective on Student Learning. Educ. Psychol. *45*, 185–202. https://doi.org/10.1080/00461520.2010.493471.
48. Schraw, G., Bruning, R., and Svoboda, C. (1995). Sources of Situational Interest. J. Read. Behav. *27*, 1–17. https://doi.org/10.1080/10862969509547866.
49. Guthrie, J.T., and Klauda, S.L. (2016). Engagement and Motivational Processes in Reading. In Handbook of individual differences in reading: Reader, text, and context, P. Afflerbach, ed. (Routledge), pp. 41–53.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
R x64 4.0.3	R Software Foundation	https://www.r-project.org/
REEMtree	REEMtree package	https://cran.r-project.org/web/packages/REEMtree/index.html

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The Program for International Student Assessment (PISA) is a project led by the Organization for Economic Co-operation and Development (OECD) that aims to assess the reading, mathematics, and science skills of 15-year-old students worldwide. In this study, 361 schools and 12,058 students from the four Chinese provinces/municipalities (Beijing, Shanghai, Jiangsu, and Zhejiang) in 2018 PISA data were used as samples.

METHOD DETAILS

The dataset for this study is derived from PISA 2018, as outlined in the "experimental model and study participant details" section. PISA provides ten plausible values (PVs) for each student's reading proficiency, with each PV representing a random sample from the distribution of reading proficiency levels. For this study, PV1 has been selected as the dependent variable, which aligns with the assessment domain of the machine learning algorithm used to evaluate students' reading abilities. After a thorough literature review, 35 predictor variables were carefully chosen to correspond with the input domain of the machine learning model. These variables cover various aspects related to the students, their families, and their schools, as detailed in Tables 1 and 2.

The statistical model employed is based on the work of Sela and Simonoff.[32]

QUANTIFICATION AND STATISTICAL ANALYSIS

We utilized R for statistical analysis. First, we conducted descriptive statistics on the covariates, as shown in Table 4. Next, we examined the differences in reading performance across different school types using one-way ANOVA, as presented in Table 5, which indicated significant inter-school variability and underscored the necessity of employing a RE-EM model. Finally, we constructed the statistical model used in this study with the "REEMtree" package in R, and demonstrated its superior predictive performance through a comparison of RMSE metrics.