Education and Information Technologies
https://doi.org/10.1007/s10639-024-13110-2

Stacking: An ensemble learning approach to predict student performance in PISA 2022

Ersoy Öz[1] · Okan Bulut[2] · Zuhal Fatma Cellat[1] · Hülya Yürekli[1]

Received: 23 April 2024 / Accepted: 11 October 2024

Abstract

Predicting student performance in international large-scale assessments (ILSAs) is crucial for understanding educational outcomes on a global scale. ILSAs, such as the Program for International Student Assessment and the Trends in International Mathematics and Science Study, serve as vital tools for policymakers, educators, and researchers to examine the effectiveness of educational systems worldwide. By accurately predicting student performance, policymakers and educators can identify trends, disparities, and areas for improvement in educational practices and policies. Researchers can utilize traditional statistical methods or machine learning algorithms to predict student achievement, enabling proactive interventions to support struggling learners and enhance overall educational quality. In this study, we employed stacking, an ensemble machine-learning algorithm, to predict student performances in large-scale assessments based on a wide range of predictors. Without filling in missing data or categorizing the outcome variable, we predicted student performances using the stacking method and then compared the results with those generated by three boosting algorithms and blending. Our findings revealed that stacking outperformed the boosting and blending methods, yielding more stable and accurate predictions. Our analysis encompassed the 80 countries that participated in the administration of PISA 2022. Compared to the three boosting algorithms and blending, we found that stacking demonstrated superior performance with the lowest error metrics for most countries. Robust linear mixed-effects models also indicated that stacking produced significantly lower MAPE, MAE, and MSE values than boosting and blending. Overall, our findings emphasize that stacking is one of the most accurate methods to predict student performance in large-scale assessments.

Keywords PISA · Machine learning · Boosting · Stacking · Blending · Ensemble learning

Extended author information available on the last page of the article

Published online: 29 October 2024

1 Introduction

In 1960, the inaugural international large-scale achievement testing coordinated and sponsored by the Governing Board of the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute for Education marked a pivotal moment. This historic event aimed to assess the academic proficiency of 13-year-old students across twelve nations, including Belgium, England, Finland, France, the Federal Republic of Germany, Israel, Poland, Scotland, Sweden, Switzerland, the USA, and Yugoslavia (Foshay et al., 1962). Following this landmark endeavor, the landscape of international large-scale assessments (ILSAs) has evolved significantly. The International Association for the Evaluation of Educational Achievement (IEA) pioneered continuous assessment cycles with the inception of the Trends in Mathematics and Science Study (TIMSS) in 1995, complemented by the Progress in Reading Literacy Study (PIRLS) in 2001. Concurrently, the Organisation for Economic Co-operation and Development (OECD) initiated the Programme for International Student Assessment (PISA) cycle of studies in the late 1990s (Kirsch et al., 2013).

ILSAs offer a multitude of benefits, spanning from the provision of extensive international assessment datasets accessible for secondary analysis by scholars to serving as benchmarks for scrutinizing trends in academic accomplishment within participating countries (Johansson, 2016). Through comprehensive databases, including indicators of student performance across diverse domains, as well as variables related to student, household, school, and teacher demographics, researchers can predict academic attainment and explore the significant factors influencing student success. Hence, ILSAs have become instrumental in facilitating cross-country comparisons across a spectrum of educational dimensions. Their datasets encapsulate a myriad of potential outcomes, encompassing academic achievement measures and, at times, motivations, attitudes, skills, perceptions of the quality of school life, and anticipated behaviors—often termed "non-cognitive" attributes (Torney-Purta & Amadeo, 2013).

To date, educational researchers have employed diverse methodologies for secondary analyses of ILSA data. For instance, scholars have investigated the interconnections among various factors, such as students' non-cognitive attributes, interactions with information and communication technology, levels of anxiety, self-concept, instrumental motivation, academic performance, engagement, self-confidence, self-perceptions, teacher efficacy, teaching practices, and school climate, among others (e.g., Chen & Hastedt, 2022; D'Agostino et al., 2022; Gao, 2014; Kahraman, 2014; Miscevic-Kadijevic, 2015; Nilsen & Teig, 2022; Örnek et al., 2023; Shen, 2010; Shi, 2014; Yu, 2012).

In addition to correlational analyses, scholars have also explored variations in student achievement both across and within countries by employing a diverse array of analytical methods. These include regression analysis (Bidegain & Mujika, 2020; Perry et al., 2022), multivariate analysis of variance (Forbes et al., 2020; Rodríguez et al., 2020), hierarchical linear modeling (Bhutoria & Aljabri, 2022; Chen & Cui, 2020), and structural equation modeling (Khine et al., 2020;

Stadler, 2020). Furthermore, some researchers have leveraged educational data mining techniques to predict student performance, identify undesirable behaviors, and categorize student profiles (e.g., Alshareef et al., 2020).

The accurate prediction of student performance holds significant importance in all levels of education, prompting extensive research efforts. Predicting student performance can ultimately lead to the development of more effective strategies for fostering academic success. Through accurate predictions of student performance in ILSAs, educators and policymakers can tailor educational strategies to address specific needs and improve overall learning outcomes on a global scale. In recent years, the integration of machine learning techniques has surged in its prominence for predicting students' academic trajectories (Alam & Mohanty, 2022). Decision trees (DTs), artificial neural networks (NNs), support vector machines (SVMs), random forests (RFs), naïve Bayes (NB), *k*-nearest neighbors (*k*NNs), and logistic regression (LR) have emerged as leading algorithms and methodologies in educational data mining, adept at predicting student performance (Batool et al., 2023; Xiao et al., 2022). However, these methodologies still fall short in capturing the complex interplay among variables, particularly in datasets characterized by a large number of highly correlated variables (Guo et al., 2015). Hence, more robust methods are necessary for building accurate predictive models utilizing large databases from ILSAs.

In this study, we utilize stacking–an ensemble learning approach in machine learning–as a novel avenue for predicting student performance in the context of ILSAs. The process of stacking involves training a learner to merge the individual learners (Zhou, 2012). While stacking has been utilized in different educational research contexts, such as predicting academic achievement (Chanamarn et al., 2016) and anxiety levels (Daza et al., 2023) among university students, as well as detecting cheating in large-scale assessments (Zhou & Jiao, 2023), its potential in predicting student performance in ILSAs has not been explored. This research involved not only the implementation of stacking but also the evaluation of its performance compared to four other ensemble learning techniques. These techniques include three distinct boosting algorithms—Extreme Gradient Boosting (XGBoost), Histogram-based Gradient Boosting (HGB), and Light Gradient Boosting (LightGBM)—and blending. Among these, XGBoost and LightGBM have been commonly employed for predicting student performance in ILSAs (e.g., Acıslı-Celik & Yeşilkanat, 2023; Cao et al., 2024; Lee & Lee, 2021; Liu et al., 2022). However, the prediction of student performance in ILSAs using HGB and blending techniques remains unexplored to our knowledge. Thus, this study is the first exploration of using stacking, HGB, and blending algorithms to predict the performance of students involved in ILSAs. By leveraging the stacking method, we aim to demonstrate how to enhance the accuracy and robustness of predictive models, which will enrich our understanding of the utility of ILSAs in predicting academic achievement.

2 Literature review

A substantial body of research in educational data mining has been dedicated to predicting students' academic performance in ILSAs using machine-learning methods (Chen & Zhai, 2023). Baker (2010) categorized educational data mining methods into several distinct groups: Prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. Within the context of prediction, there are three main sub-branches: (1) classification (for binary or categorical target variables), (2) regression (for continuous target variables), and (3) density estimation (i.e., the target variable is a probability density function) (Baker, 2010; Romero & Ventura, 2013).

Several studies have employed classification techniques to predict student achievement in ILSAs, typically categorizing performance as either successful or unsuccessful, or as pass or fail. For example, Büyükkıdık et al. (2018) classified students based on their scientific literacy scores from PISA 2015, utilizing multilayer perceptron, LR, and SVM methods. Their results highlighted that SVM generated the highest accuracy across all tested conditions with tenfold cross-validation. Bayirli et al. (2023) classified students' mathematics achievement in PISA 2018, using RF, LR, and SVM models. According to their findings, the SVM model outperformed the rest of the models based on the classification performance measures. Koyuncu and Gelbal (2020) investigated mathematics performance in PISA 2012, employing NB, *k*NN, NN, and LR. They discovered variations in performance across different sample sizes and test data rates. Büyükkıdık (2023) also examined mathematical literacy among students using PISA 2015, employing the multilayer perceptron, J48, SVM, and NB methods. The results obtained from this study indicated that while all of these approaches were able to sufficiently categorize student achievement, none of them emerged as universally superior across all evaluation criteria. Despite the encouraging outcomes observed in achievement-focused predictive studies rooted in ILSAs, one of the limitations is the tendency of most researchers to operationalize achievement variables as categorical, even though the majority of outcome variables in ILSAs (e.g., test scores representing student performance) are continuous. This practice often leads to a significant loss of valuable information, potentially compromising the overall effectiveness of the research (Altman & Royston, 2006; Carazo-Díaz & Prieto-Valiente, 2024).

Various analytical methods have been utilized to predict students' academic success. Although traditional statistical techniques have been commonly applied in this regard, their drawback is the inadequate performance in accurately predicting students' academic performance. Researchers employing such techniques may encounter challenges such as the risk of overfitting models, the difficulty of efficiently managing large numbers of participants and predictors, and the inability to discern potential nonlinear relationships; however, regression-based machine learning methods offer promising avenues to address these issues, producing accurate models that remain highly interpretable (Puah, 2020). For instance, in a recent study focusing on immigrant students who participated in PISA 2018

(Jeganathan et al., 2022), students' academic performance was predicted using several machine learning algorithms, such as linear regression, SVM, ridge regression, RF, and XGBoost. Notably, the XGBoost technique demonstrated comparable performance, particularly when fine-tuned using hyperparameters. In another study conducted by Acıslı-Celik and Yesilkanat (2023), students' PISA 2018 science achievement scores were predicted using multiple linear regression, support vector regression, RF, and XGBoost machine-learning algorithms. The findings indicated that the XGBoost machine learning algorithm produces more accurate predictions.

Although traditional machine learning algorithms have made significant strides in knowledge discovery, they often grapple with challenges when handling complex data, including imbalanced, high-dimensional, and noisy datasets, especially when capturing the nuanced characteristics and underlying structures inherent in the data (Dong et al., 2020). Meanwhile, ensemble-learning models offer a promising solution by combining diverse algorithms, enhancing algorithmic accuracy and model stability (Dineva & Atanasova, 2020). Through this combination, ensemble models yield an optimal predictive model that surpasses the performance of individual base algorithms (Dineva & Atanasova, 2020). In addition, ensemble learning methods effectively tackle issues such as class imbalance, concept drift, and the curse of dimensionality (Sagi & Rokach, 2018). Recent studies suggest that ensemble-learning methods outperform individual machine learning techniques in predicting student performance (Teoh et al., 2022). Hence, the primary objective of this study is to harness stacking—an ensemble method capable of handling missing data and suitable for continuous outcome variables—to predict student performance and evaluate its effectiveness compared to boosting and blending algorithms.

3 Methods

3.1 Dataset

The dataset utilized in this study came from the 2022 administration of PISA, a renowned ILSA program designed to assess the proficiencies of 15-year-old students in reading, mathematics, and science. Conducted triennially, PISA systematically compares and scrutinizes students' academic performance across diverse educational systems worldwide. Comprising extensive and diverse student performance data from various global educational systems, the PISA dataset represents a robust and invaluable resource for machine learning algorithms. These algorithms can scrutinize student achievements across multiple dimensions to tailor learning processes and instructional designs.

In this study, we employed various machine-learning algorithms to predict student performance in the subject areas of mathematics, reading, and science. We utilized data from all countries participating in the PISA 2022 administration. The PISA 2022 dataset consisted of the full set of responses from individual students (i.e., the student questionnaire file), as well as students' academic performance in mathematics, reading, and science, each featuring ten plausible values (PVs). For each subject area (i.e., mathematics, reading, and science), we eliminated irrelevant variables (e.g., student and school IDs), the student birth-month variable, country-specific items, variables that were not subject-specific, and variables with 100% missing values. This procedure yielded a separate dataset for each subject area with the remaining variables in the PISA 2022 dataset.

3.2 Machine learning algorithms

This study utilizes three ensemble learning techniques (i.e., stacking, blending, and boosting) to predict student performances in reading, mathematics, and science. Stacking and blending involves training multiple diverse base models in parallel and then using a meta-model to combine their predictions for improved results. Unlike stacking and blending, boosting is an iterative process where a series of weak learners (i.e., simple models) are trained sequentially. Each new model focuses on correcting the errors made by the previous ones, gradually improving the ensemble's overall performance. The following sections describe these techniques in more detail.

3.2.1 Stacking

Stacking, developed by Wolpert in 1992, is an ensemble learning technique that combines several prediction models and uses their outputs as inputs for a next-level model (meta-model), with the goal of improving overall prediction performance (Breiman, 1996; Smyth & Wolpert, 1997). The process works by "stacking" the forecasts from several models and using this information to train a new model (Wolpert, 1992). Carefully choosing the base classifier in a stacked generalization ensemble classifier reduces the rate of misclassification. By further training the base classifier's predicted decision to create a meta-classifier model, the best base classifiers are chosen. The effectiveness of the meta-learner model, in contrast to base learning, can be evaluated by introducing noise into the raw dataset. For increased prediction accuracy, these stacked generalization classifier features have been applied in a number of research domains (Manav-Demir et al., 2024). A meta-learner decreases bias and increases flexibility (Manav-Demir et al., 2024).

Stacking is a generalization process based on multiple, small models (Fig. 1). The original learning set itself is referred to as level 0. At this stage, the original learning data set is the direct operating environment for the model or models (i.e., generalizers, base learners). These base learners create predictions using the data set's raw data, which are subsequently recorded for usage at a higher level. Different machine learning techniques can be used to create models at level 0, and each model can produce a unique prediction. Level 0 learners could use various machine learning models like DTs, NNs, SVMs, and *k*NNs. The predictions (i.e., the output) of these learners are then used as the input for the "level 1" learner, which aims to combine these predictions in a meaningful way to improve the overall prediction (Wolpert, 1992).

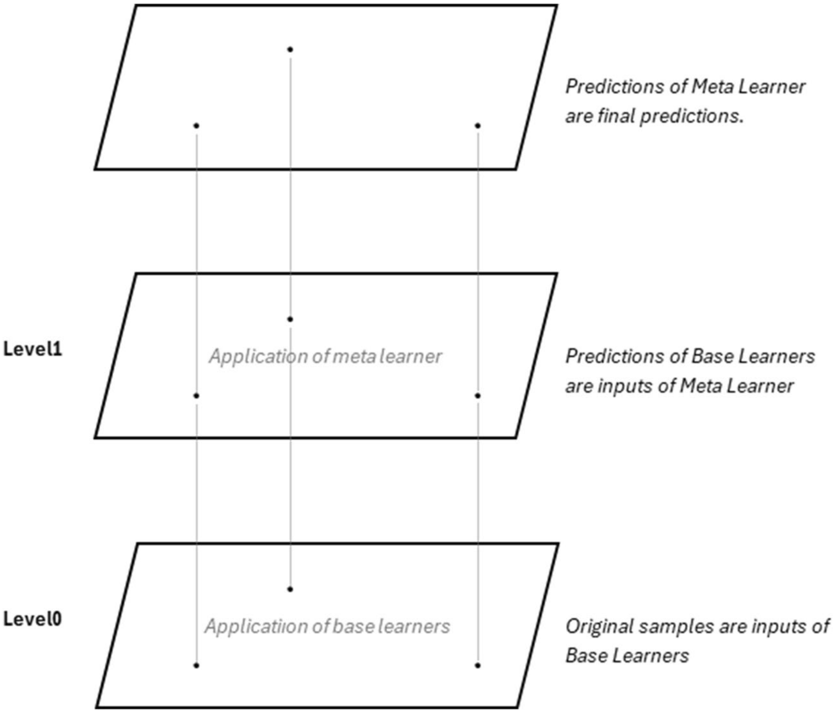
Predictions of Meta Learner are final predictions.
Level1
Application of meta learner
Predictions of Base Learners are inputs of Meta Learner
Level0
Application of base learners
Original samples are inputs of Base Learners

Fig. 1 An illustration of the stacking method

The level 0 base learners' predictions are combined and utilized as inputs in level 1, which is the subsequent stage (Fig. 1). In this case, the level 0 predictions serve as training data for a new "meta" model. This level typically consists of a model (meta-learner) that aggregates predictions from various level 0 models and produces a final prediction or decision based on these predictions. In level 1, predictions combined in this manner are analyzed using a more sophisticated model, yielding the ultimate prediction. In the stacking method, the selection of the level 1 model may vary depending on the type of basic models and the characteristics of the problem (Wolpert, 1992). Some popular algorithms that can be used at level 1 are LR, linear regression, ridge regression, lasso regression, DTs, and RF. In this study, the ridge regression technique was used to prevent overfitting and to increase the generalization ability of the model (Cui et al., 2021). Ridge regression is a regularized version of linear regression, which is used to avoid overfitting. By performing ridge regression, it is prevented from the model fitting the training set too closely, which may result in subpar performance on fresh, untested data. The "alpha" parameter in ridge regression is used to establish the degree of regularization, or to see how "tight" the model will be. Less regularization is indicated by a low alpha value. As a result, the model can fit the training set of data more closely by allowing the coefficients to increase in value; however, overfitting may result from this process. Greater regularization is correlated with a higher alpha value, leading to smaller coefficients and improving the model's capacity for generalization. Underfitting can occur when an alpha value is too high because it hinders the model's ability to fit the training set (Hoerl & Kennard, 1970).

The two stages of the stacked generalization method are as follows: at level 0, different models generate distinct predictions, and at level 1, these predictions are combined to create a more comprehensive prediction. By combining the advantages of several models, this technique seeks to increase overall prediction success (Wolpert, 1992).

3.2.2 Boosting

Given that the dataset utilized in this investigation comprises of missing data, it is imperative to employ base learners that possess the capability to handle missing data (Chen & Guestrin, 2016; Ke et al., 2017; Nair et al., 2023). Thus, we have used three different boosting algorithms for this study: XGBoost, HGB, and LightGBM.

XGBoost algorithm Originally developed and released by Chen and Guestrin in 2016, XGBoost is a large-scale machine-learning algorithm that can learn from tree-based models through boosting. XGBoost uses a regularized learning objective to train its model. To forecast the result based on the samples and features in the dataset, XGBoost uses a number of additive functions. These functions keep the model from becoming too complex, which prevents overfitting and helps to identify more straightforward but effective predictive functions. By adding new functions that improve the model at each iteration, the model is gradually improved during the XGBoost training process. This situation frequently calls for the optimization of the objective using a second-order approximation, enabling quick and efficient model training (Chen & Guestrin, 2016). Being a supervised learning algorithm, XGBoost improves tree-based models by combining weak learners to produce a strong model. Its resilience to overfitting and insensitivity to outliers make it stand out. It can function fast and effectively even in the case of missing data. Although XGBoost can be used in many different contexts, its ability to handle varying data is limited (Manav-Demir et al., 2024).

HGB The HGB classifier, developed by Friedman in 1999, is preferred because it uses benchmark accuracy based on histogram features and has a lower computational cost than the gradient boosting classifier. A HGB classifier, in contrast to a gradient boosting classifier, uses a feature histogram to determine the best splitting feature points. Thus, the histogram data structure contributes to a reduction in computational complexity (Hossain & Deb, 2021). In HGB, the values of continuous attributes are grouped (binned), reducing data or attributes. This method classifies the numerous unique values of continuous attributes into fewer "bins", reducing the computational load and speeding up the model's training process (Raj et al., 2023).

LightGBM Microsoft first created and released LightGBM in 2017. LightGBM is an implementation of the Gradient Boosting Decision Tree (GBDT) algorithm. The purpose of this algorithm is to solve the problems arising from high dimensionality and large data sizes in big datasets. Two novel approaches are presented by LightGBM: Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS). For the purpose of calculating information gain, GOSS concentrates on data instances with larger gradients, whereas EFB bundles mutually exclusive features in sparse feature spaces to minimize the number of effective features. By utilizing these strategies, LightGBM is able to maintain accuracy while drastically accelerating the training process in comparison to conventional GBDT methods. Large-scale machine learning applications benefit greatly from LightGBM's fast training times and high accuracy rates, which make it an exceptional machine learning solution-particularly for datasets of a considerable size (Ke et al., 2017; Kunapuli, 2023).

3.2.3 (Advanced) blending

Blending and stacking are two similar techniques in ensemble learning. The primary distinction between these two ensemble learning methods lies in the fact that the base learners are not trained using predictions derived from the training set, but rather with predictions derived from the validation set, which is initially set aside at the start of the process, typically constituting 10% or 20% of the total dataset (Brownlee, 2021; Wu et al., 2021). The blending technique, by excluding a portion of the training set, offers a more efficient and expedited process compared to stacking. This method entails utilizing a reduced amount of data (approximately 90% or 80%) during the training of base learners (Wu et al., 2021).

3.3 Data analysis

Python (version 3.8.10) has been used for data preparation and running the algorithms. We started with data preparation as the first step. Next, we performed level 0 and level 1 analyses in the second and third steps, respectively. The procedure we followed during each step is illustrated in Fig. 2 and explained in detail in the following section. The stacking and blending processes are visually differentiated in Fig. 2, with green and orange colors assigned to each process.

First Step: Preparation

In the first step, original (raw) data was prepared for the application of base learners. Values in the dataset indicating "valid skip," "not applicable," "invalid," and "no response" were converted to NaN (i.e., missing). Furthermore, since the results of this study will be produced on a country-by-country basis, the dataset has been examined separately for each country. Variables that contain 100% missing values for that country were removed.

Second Step: Level 0

In the initial phase, preparations for the algorithms to be applied in the first layer (level 0) of the stacking and blending processes have been made for each of the country datasets obtained. For this purpose, the data has first been split into 80% for training and 20% for testing for the stacking process, and 60% for training, 20%

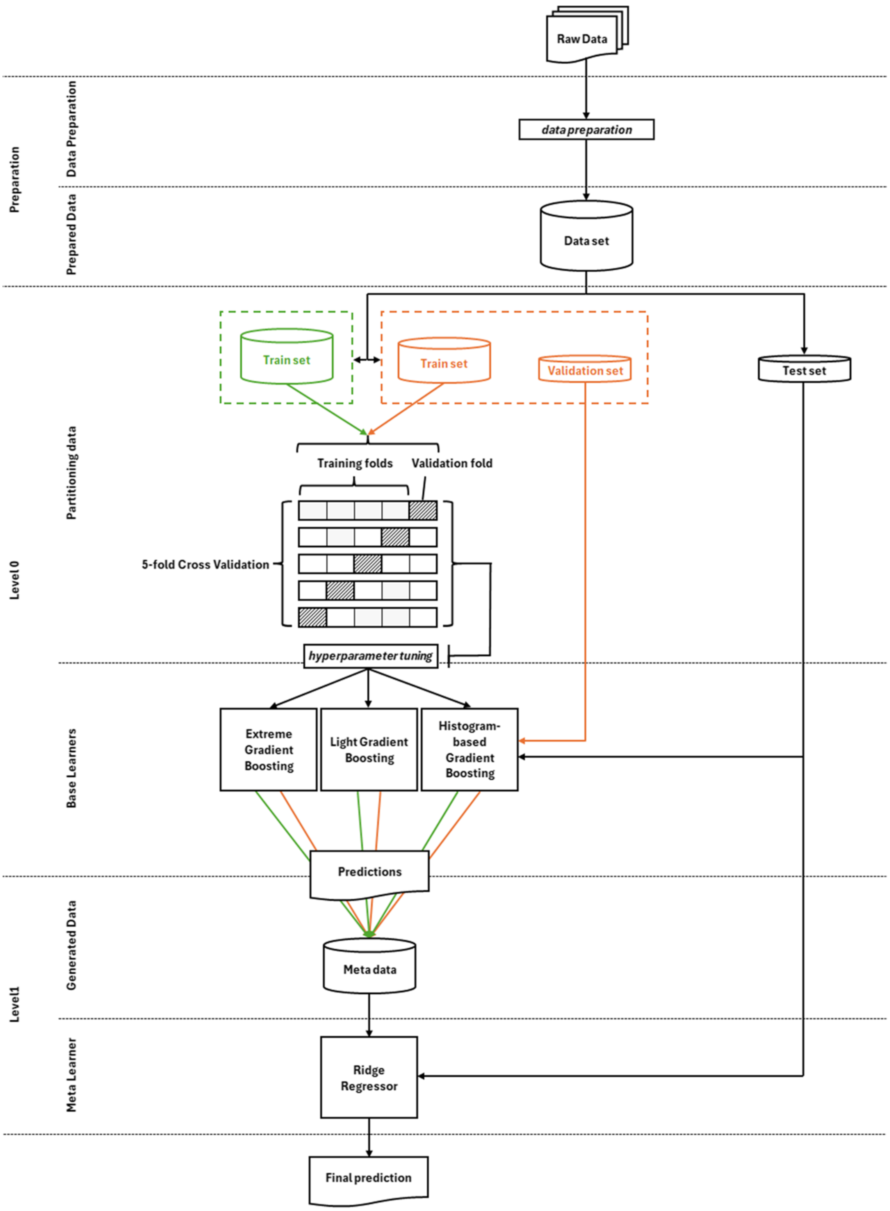

Fig. 2 Approaches utilized in the current study

for validation, and 20% for testing for the blending process. The primary goal is to make the most accurate predictions possible for the separated 20% test set, either using the training dataset in the stacking process or using the validation dataset in the blending process. To achieve this, parameter optimization has been performed for the parameters of the algorithms listed in Table 1.

Table 1 Optimized hyperparameters for base learners

Model	Parameters
XGBoost	Number of estimators: [100, 200, 300] Learning rate: [0.01, 0.1, 0.3] Maximum depth: [3, 5, 6]
HGB	Maximum iterations: [100, 200, 300] Learning rate: [0.01, 0.1, 0.2] Maximum depth: [None, 3, 5]
LightGBM	Number of estimators: [100, 200, 300] Learning rate: [0.01, 0.1, 0.2] Maximum depth: [-1, 3, 5]

During parameter optimization, a fivefold cross-validation method was utilized, and predictions were obtained for each algorithm using the optimal parameters. The default values were used for the algorithm parameters that are not listed in Table 1.

Final Step: Level 1

During the level-1 layer of the stacking process, the predictions made by the base learners (level 0) obtained with the training data are utilized as input, leading to the generation of stronger and widely applicable outcomes. The level-1 layer in the blending process receives inputs consisting of predictions obtained using the validation data at the level-0 layer. Predictions produced by each base learner are combined in this step to create a meta-model, which is selected to be the ridge regression with a regularization degree (alpha) of 1.0 (Cui et al., 2021). Making final predictions on fresh data requires the meta-model to acquire the skill of efficiently integrating the predictions of the base models. By leveraging the strengths of each model and mitigating their weaknesses, these approaches work especially well when there is diversity among modeling. More precise and trustworthy predictions are thus provided by stacking or blending than by any single base model working alone.

3.4 Model evaluation criteria

Several error metrics were calculated to evaluate the models' performance using the training and test datasets. The mean absolute percentage error (MAPE), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) metrics were employed in this study to evaluate the algorithms that produced the predicted values. The performance statistical index measures used for prediction are listed and explained in Table 2.

As the MAPE, MAE, and MSE scores exhibited non-normality, robust linear mixed-effects modeling was used to assess the significance of the observed differences in these metrics. The robust linear mixed-effects analysis was carried out using the robustlmm package (Koller, 2016) in R (R Core Team, 2023), with the dependent variables being the MAPE, MAE, and MSE scores, algorithms as fixed effects, countries as random effects, and the Stacking algorithm as the reference level.

Table 2 Model evaluation metrics

Metric	Equation	Explanation
MAPE	$\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{y_{actual}-y_{predicted}}{y_{actual}}\right\|$	Calculates the difference between actual values and predicted values as a percentage so that comparisons between problems can be made at different scales
MAE	$\frac{1}{n}\sum_{i=1}^{n}\left\|(y_{actual}-y_{predicted})\right\|$	Measures the mean size of the forecast error for a collection of forecasts
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_{actual}-y_{predicted})^2$	Displays the error measure alongside the raw data in the same entities
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{actual}-y_{predicted})^2}$	Presents the raw data and the root error measure in the same entities

Source: (Dou et al., 2020; Manav-Demir et al., 2024)

4 Results

The PISA 2022 mathematics, science, and reading performances of students from 80 countries worldwide were predicted through the application of machine learning algorithms in this study. The initial dataset was employed during the level 0 stage of the stacking procedure, utilizing XGBoost, HGB, and LightGBM algorithms as the base learners, without imputing any missing observations. For the stacking process, the predictions made by the base learners using training data at level 0 served as the inputs for the meta-learner at level 1. Predictions derived from base learners utilizing validation data at the level-0 stage were further utilized as inputs for the meta-learner in the level-1 stage during the blending procedure. The outputs of the base learners and the meta-learner were used to compute the performance metrics. This procedure was performed for each PV of each subject. The mean MAPE, MAE, MSE, and RMSE values and 95% confidence intervals were calculated separately using stacking, blending, and three boosting algorithms for each subject and all countries, and the findings were outlined in Table 3 and further explained in detail in the following section. Tables in the supplementary file list the mean MAPE, MAE, MSE, and RMSE values and 95% confidence intervals separately for each subject and all countries.

Next, a series of robust linear mixed-effects models were conducted to investigate the effect of different algorithms on mathematics, reading, and science MAPE, MAE, and MSE scores, accounting for country-level variability. The fixed effects included in the models were the five algorithms (Stacking, XGBoost, HGB, LightGBM, and Blending), with Stacking as the reference category. The random effect was specified for the countries. Table 4 shows the results of the robust linear mixed-effects models for each performance metric.

4.1 MAPE values

A total of 50 MAPE values were calculated for every subject and all countries by employing five distinct algorithms: Stacking, Blending, XGBoost, HGB, and

Table 3 The Number (%) of the countries exhibiting the lowest error values generated by each algorithm for all subjects

Subject		Mathematics	Reading	Science
Metrics	Algorithm	Number (%) of the countries	Number (%) of the countries	Number (%) of the countries
Mean MAPE	Stacking	72 (90.00)	64 (80.00)	65 (81.25)
	Blending	4 (5.00)	4 (5.00)	6 (7.50)
	XGBoost	-	1 (1.25)	1 (1.25)
	LightGBM	2 (2.50)	6 (7.50)	3 (3.75)
	Blending & Stacking	-	2 (2.50)	2 (2.50)
	Stacking & XGBoost	-	-	2 (2.50)
	Blending & LightGBM	-	-	1 (1.25)
	LightGBM & XGBoost	2 (2.50)	1 (1.25)	-
	HGB & LightGBM	-	1 (1.25)	-
	XGBoost & HGB & LightGBM	-	1 (1.25)	-
Mean MAE	Stacking	74 (92.50)	69 (86.25)	71 (88.75)
	Blending	1 (1.25)	1 (1.25)	2 (2.50)
	XGBoost	1 (1.25)	2 (2.50)	3 (3.75)
	HGB	1 (1.25)	1 (1.25)	-
	LightGBM	3 (3.75)	7 (8.75)	4 (5.00)
Mean MSE	Stacking	75 (93.75)	70 (87.50)	71 (88.75)
	Blending	1 (1.25)	1 (1.25)	2 (2.50)
	XGBoost	-	3 (3.75)	1 (1.25)
	HGB	-	1 (1.25)	-
	LightGBM	4 (5.00)	5 (6.25)	6 (7.50)

LightGBM across ten different PV values. The mean MAPE value was computed for each algorithm, subject, and country by averaging 10 MAPE values calculated for each PV of that subject in that country. In addition, 95% confidence intervals for the mean MAPE values were calculated (Please see supplementary file Tables S1.1, S1.5, and S1.9). Furthermore, robust linear mixed-effects models were performed to examine the effect of different algorithms on the MAPE scores for each subject, accounting for country-level variability.

Figure 3 displays the algorithms with the lowest mean MAPE values for all countries separately for mathematics, reading, and science. Among 80 nations where algorithms were performed, stacking produced the lowest mean MAPE values for 72 nations (90%) in mathematics, 64 nations (80%) in reading, and 65 nations (81.25%) in science (Table 3). This implies that the student performance predictions generated by stacking are, on average, closely align with the actual student performances in at least 80% of cases.

The examination of mean MAPE values for mathematics across 80 nations revealed that LightGBM generated the lowest mean MAPE value for two nations

Table 4 Robust linear mixed-effects model results

Criteria	Subject	Fixed effects	Estimate	Std. Error	*t*-value
MAPE	Mathematics	Intercept	0.10	0.00	76.65
		XGBoost	0.00	0.00	13.79
		HGB	0.00	0.00	13.76
		LightGBM	0.00	0.00	12.63
		Blending	0.00	0.00	23.22
	Reading	Intercept	0.12	0.00	68.70
		XGBoost	0.00	0.00	10.60
		HGB	0.00	0.00	12.86
		LightGBM	0.00	0.00	11.00
		Blending	0.00	0.00	21.25
	Science	Intercept	0.11	0.00	76.75
		XGBoost	0.00	0.00	11.51
		HGB	0.00	0.00	11.95
		LightGBM	0.00	0.00	10.74
		Blending	0.00	0.00	20.00
MAE	Mathematics	Intercept	41.66	0.74	56.33
		XGBoost	0.19	0.01	14.15
		HGB	0.18	0.01	13.90
		LightGBM	0.18	0.01	13.45
		Blending	0.37	0.01	28.10
	Reading	Intercept	47.54	0.83	57.23
		XGBoost	0.16	0.01	12.21
		HGB	0.19	0.01	14.97
		LightGBM	0.17	0.01	13.61
		Blending	0.35	0.01	26.98
	Science	Intercept	46.26	0.75	61.91
		XGBoost	0.17	0.02	10.87
		HGB	0.17	0.02	11.43
		LightGBM	0.16	0.02	10.13
		Blending	0.39	0.02	25.64
MSE	Mathematics	Intercept	2794.12	98.46	28.38
		XGBoost	23.15	1.72	13.50
		HGB	24.96	1.72	14.55
		LightGBM	23.26	1.72	13.56
		Blending	48.56	1.72	28.31
	Reading	Intercept	3623.65	124.95	29.00
		XGBoost	23.95	2.02	11.87
		HGB	28.65	2.02	14.20
		LightGBM	25.60	2.02	12.69
		Blending	53.28	2.02	26.40
	Science	Intercept	3414.42	109.05	31.31
		XGBoost	23.42	2.24	10.45

Table 4 (continued)

Criteria	Subject	Fixed effects	Estimate	Std. Error	*t*-value
		HGB	25.87	2.24	11.54
		LightGBM	22.92	2.24	10.22
		Blending	57.47	2.24	25.63

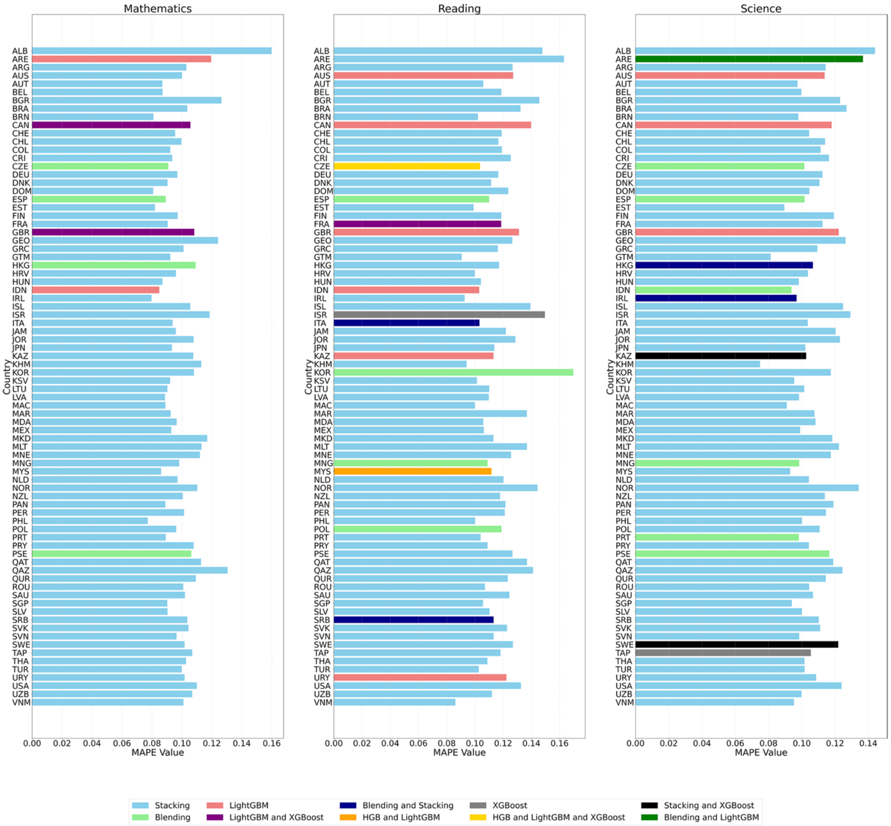

Fig. 3 The mean MAPE values for mathematics, reading, and science

(2.5%), whereas Blending achieved the lowest mean MAPE value for four nations (5%). Furthermore, XGBoost and LightGBM shared the same lowest mean MAPE value for two nations (2.50%). Noteworthy is the fact that stacking yielded the lowest mean MAPE values for 72 nations (90.00%) in the domain of mathematics. The mean MAPE values for HGB were greater than those of the other four algorithms in mathematics across all countries. According to the robust linear mixed-effects model results, the intercept was significant ($\beta = 0.10$, $SE = 0.00$, $t = 76.65$), representing the mean mathematics MAPE score for the Stacking. The effects of XGBoost ($\beta = 0.00$, *SE* = 0.00, *t* = 13.79), HGB (β = 0.00, *SE* = 0.00, *t* = 13.76), LightGBM (β = 0.00, *SE* = 0.00, *t* = 12.63), and Blending (β = 0.00, *SE* = 0.00, *t* = 23.22) were all significant, indicating increases in mathematics MAPE scores compared to the Stacking. The random effects showed variability across countries (*SD* = 0.01).

Across the 80 nations included in our analysis of mean MAPE values for reading, Blending achieved the lowest mean MAPE value for four nations (5.00%). XGBoost delivered the lowest mean MAPE value for one nation (1.25%), while LightGBM produced the lowest mean MAPE value for six nations (7.50%). Additionally, Blending and Stacking both displayed the same lowest mean MAPE value for two nations (2.50%). Xgboost and LightGBM both achieved the same lowest mean MAPE value for one nation (1.25%), HGB and LightGBM both achieved the same lowest mean MAPE value for one nation (1.25%), and HGB, LightGBM, and XGBoost attained the same lowest mean MAPE value for one nation (1.25%). It is important to highlight that stacking produced the lowest mean MAPE values for 64 nations (80.00%) within the field of reading. The findings of the robust linear mixed-effects model indicated that the intercept was significant (β = 0.12, *SE* = 0.00, *t* = 68.70), representing the mean value of the dependent variable for the Stacking. The effects of XGBoost (β = 0.00, SE = 0.00, t = 10.60), HGB (β = 0.00, *SE* = 0.00, *t* = 12.86), LightGBM (β = 0.00, *SE* = 0.00, *t* = 11.00), and Blending (β = 0.00, *SE* = 0.00, *t* = 21.25) were all significant, indicating increases in the reading MAPE scores compared to the Stacking. The random effects showed variability across countries (*SD* = 0.02).

In our analysis of mean MAPE values for science across 80 nations, Blending achieved the lowest mean MAPE value for six nations (7.5%). XGBoost demonstrated the lowest mean MAPE value for one nation (1.25%), while LightGBM showcased the lowest mean MAPE value for three nations (3.75%). Moreover, Blending and Stacking produced the same lowest mean MAPE value for two nations (2.50%). XGBoost and Stacking exhibited identical lowest mean MAPE values for two nations (2.50%), and Blending and LightGBM exhibited identical lowest mean MAPE values for one nation (1.25%). Notably, stacking emerged as the method with the lowest mean MAPE values for 65 nations (81.25%) within the realm of science. The findings of the robust linear mixed-effects model demonstrated that the intercept was significant (β = 0.11, *SE* = 0.00, *t* = 76.75), representing the mean science MAPE score for the Stacking. The effects of XGBoost (β = 0.00, *SE* = 0.00, *t* = 11.51), HGB (β = 0.00, *SE* = 0.00, *t* = 11.95), LightGBM (β = 0.00, *SE* = 0.00, *t* = 10.74), and Blending (β = 0.00, *SE* = 0.00, *t* = 20.00) were all significant, indicating increases in science MAPE scores compared to the Stacking. The random effects showed variability across countries (*SD* = 0.01).

4.2 MAE values

50 MAE values were computed for each subject and country using five different algorithms (i.e., stacking, blending, XGBoost, HGB, and LightGBM) for ten different PV values. The mean MAE value was determined for each algorithm, subject, and country by averaging the 10 MAE values calculated for each PV of that subject

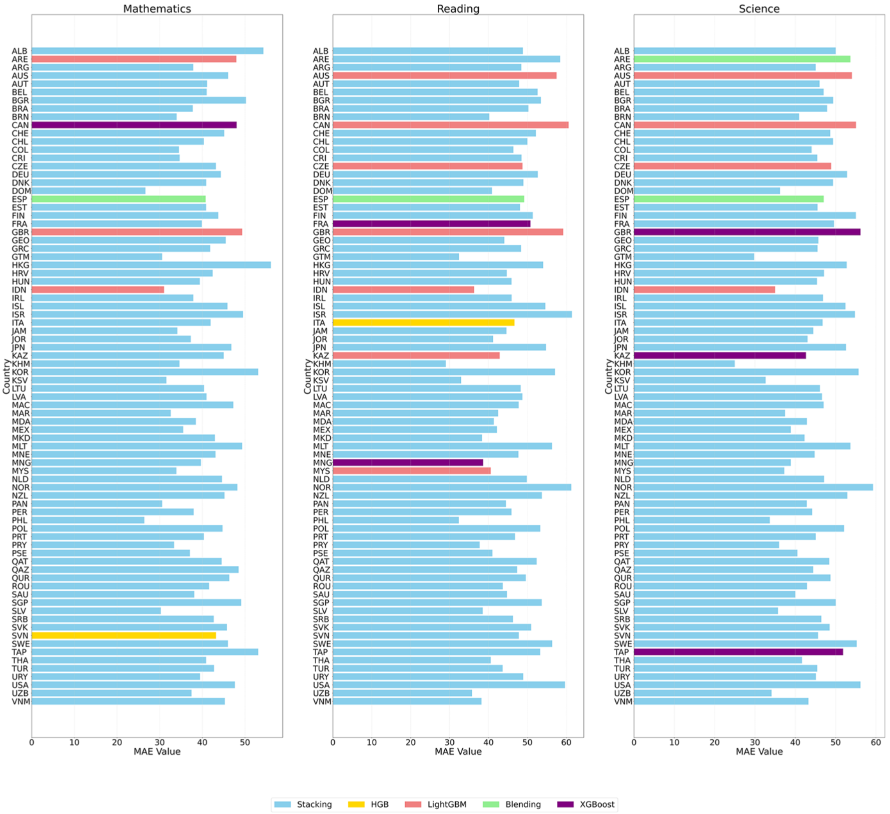

Fig. 4 The mean MAE values for mathematics, reading and science

in that country. Furthermore, 95% confidence intervals for the mean MAE values were calculated (Please refer to supplementary file Tables S1.2, S1.6, and S1.10). Moreover, robust linear mixed-effects models were performed to examine the effect of different algorithms on the MAE scores for each subject, accounting for country-level variability.

Figure 4 illustrates the algorithms that exhibited the lowest mean MAE values for each country individually in mathematics, reading, and science. Out of the 80 countries where algorithms were applied, stacking yielded the lowest mean MAE values for 74 countries (92.50%) in mathematics, 69 countries (86.25%) in reading, and 71 countries (88.75%) in science (see Table 3).

Among the 80 nations analyzed for mean MAE values in mathematics, blending produced the lowest mean MAPE value for one nation (1.25%). In addition, XGBoost yielded the lowest mean MAE value for one nation (1.25%), while LightGBM achieved the lowest mean MAE value for three nations (3.75%), and HGB produced the lowest mean MAE value for one nation (1.25%). Stacking, however, yielded the lowest mean MAE values for 74 nations (92.50%) in mathematics.

Consequently, the lower MAE values observed in stacking across various countries indicate that the student performance predictions generated through stacking are more closely aligned with the real student performances. The findings of the robust linear mixed-effects model demonstrated that the intercept was significant ($\beta=41.66$, $SE=0.74$, $t=56.33$), representing the mean mathematics MAE score for the Stacking. The effects of XGBoost ($\beta=0.19$, $SE=0.01$, $t=14.15$), HGB ($\beta=0.18$, $SE=0.01$, $t=13.90$), LightGBM ($\beta=0.18$, $SE=0.01$, $t=13.45$), and Blending ($\beta=0.37$, $SE=0.01$, $t=28.10$) were all significant, indicating increases in mathematics MAE compared to the Stacking. The random effects showed variability across countries (variance = 41.59, $SD=6.45$).

Of the 80 nations considered in our study of mean MAE values in reading, Blending produced the lowest mean MAPE value for one nation (1.25%). Furthermore, XGBoost yielded the lowest mean MAE value for two nations (2.50%), LightGBM obtained the lowest mean MAE value for seven nations (8.75%), and HGB produced the lowest mean MAE value for one nation (1.25%). On the other hand, stacking produced the lowest mean MAE value for 69 nations (86.25%). The findings of the robust linear mixed-effects model indicated that the intercept was significant ($\beta=47.54$, $SE=0.83$, $t=57.23$), representing the mean reading MAE score for the Stacking. The effects of XGBoost ($\beta=0.16$, $SE=0.01$, $t=12.21$), HGB ($\beta=0.19$, $SE=0.01$, $t=14.97$), LightGBM ($\beta=0.17$, $SE=0.01$, $t=13.61$), and Blending ($\beta=0.35$, $SE=0.01$, $t=26.98$) were all significant, indicating increases in reading MAE compared to the Stacking. The random effects showed variability across countries (variance = 52.47, $SD=7.24$).

In our analysis of mean MAE values in science for 80 nations, Blending produced the lowest mean MAPE value for two nations (2.5%). Additionally, XGBoost demonstrated the lowest mean MAE value for three nations (3.75%). LightGBM achieved the lowest mean MAE value for four nations (5%). The mean MAE values for HGB were higher compared to the other three algorithms across all countries. Stacking outperformed the other algorithms by yielding the lowest mean MAE value for 71 nations (86.25%). The outcomes of the robust linear mixed-effects model demonstrated that the intercept was significant ($\beta=46.26$, $SE=0.75$, $t=61.91$), representing the mean science MAE score for the Stacking. The effects of XGBoost ($\beta=0.17$, $SE=0.02$, $t=10.87$), HGB ($\beta=0.17$, $SE=0.02$, $t=11.43$), LightGBM ($\beta=0.16$, $SE=0.02$, $t=10.13$), and Blending ($\beta=0.39$, $SE=0.02$, $t=25.64$) were all significant, indicating increases in science MAE compared to the Stacking. The random effects showed variability across countries (variance = 42.45, $SD=6.52$).

4.3 MSE values

For each subject and country, a total of 50 MSE values were computed using five different algorithms – Stacking, Blending, XGBoost, HGB, and LightGBM—across 10 different PV values. The mean MSE value was established for each algorithm, subject, and country by averaging the 10 MSE values calculated for each PV of that subject in that country. Moreover, 95% confidence intervals for the mean MSE values were calculated (Provided in supplementary file Tables S1.3, S1.7, and S1.11).

Further, robust linear mixed-effects models were performed to examine the effect of different algorithms on the MSE scores for each subject, accounting for country-level variability.

The algorithms with the lowest mean MSE values for mathematics, reading, and science for each country are presented in Fig. 5. Stacking yielded the lowest mean MSE values for 75 countries (93.75%) in mathematics, 70 countries (87.50%) in reading, and 71 countries (88.75%) in science out of the 80 countries where algorithms were applied (Table 3). Therefore, it can be observed that in most countries, stacking predictions closely correspond to actual student performances (Table 3).

The MSE values in mathematics were analyzed across 80 nations. Blending producd the lowest mean MAPE value for one nation (1.25%), whereas LightGBM produced the lowest mean MSE values for four nations (5.00%). Comparatively, the mean MSE values generated by HGB and XGBoost were higher than those of the other three algorithms. Notably, stacking exhibited the lowest mean MSE values for most nations, precisely 75 nations (93.75%), in mathematics. The findings of the robust linear mixed-effects model demonstrated that the intercept was significant

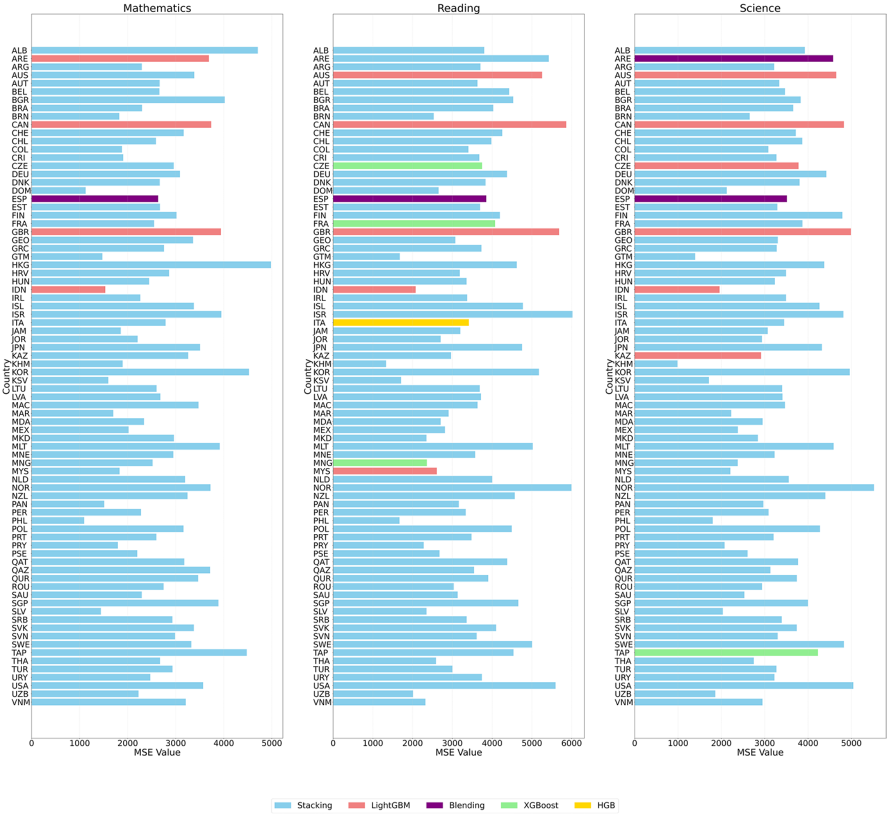

Fig. 5 The Mean MSE Values for mathematics, reading, and science

($\beta = 2794.12$, $SE = 98.46$, $t = 28.38$), representing the mean mathematics MSE score for the Stacking. The effects of XGBoost ($\beta = 23.15$, $SE = 1.72$, $t = 13.50$), HGB ($\beta = 24.96$, $SE = 1.72$, $t = 14.55$), LightGBM ($\beta = 23.26$, $SE = 1.72$, $t = 13.56$), and Blending ($\beta = 48.56$, $SE = 1.72$, $t = 28.31$) were all significant, indicating increases in mathematics MSE compared to the Stacking. The random effects showed substantial variability across countries (variance = 737,180.10, $SD = 858.59$).

According to the analysis of mean MSE values in reading across 80 countries, Blending attained the lowest mean MAPE value for one nation (1.25%), and XGBoost showcased the lowest mean MSE value for three nations (3.75%). Conversely, LightGBM demonstrated the lowest mean MSE values for six nations (7.50%), while HGB exhibited the lowest mean MSE values for one nation (1.25%). It is worth mentioning that stacking emerged as the frontrunner in terms of the lowest mean MSE values for a substantial majority of nations, precisely 70 nations (87.50%), within the domain of reading. The findings of the robust linear mixed-effects model revealed that the intercept was significant ($\beta = 3623.65$, $SE = 124.95$, $t = 29.00$), representing the mean reading MSE score for the Stacking. The effects of XGBoost ($\beta = 23.95$, $SE = 2.02$, $t = 11.87$), HGB ($\beta = 28.65$, $SE = 2.02$, $t = 14.20$), LightGBM ($\beta = 25.60$, $SE = 2.02$, $t = 12.69$), and Blending ($\beta = 53.28$, $SE = 2.02$, $t = 26.40$) were all significant, indicating increases in reading MSE compared to the reference algorithm. The random effects showed substantial variability across countries (variance = 1,187,180.90, $SD = 1089.58$).

Upon analyzing the mean MSE values in science across 80 countries, it was observed that XGBoost had the lowest mean MSE value for one nation (1.25%). Blending also produced the lowest mean MAPE value for two nations (2.5%). LightGBM, on the other hand, demonstrated the lowest mean MSE values for six nations (7.5%). HGB, however, yielded higher mean MSE values than the other three algorithms. It is important to note that Stacking emerged as the top performer in achieving the lowest mean MSE values for a vast majority of nations, precisely 71 nations (88.75%), within the science field. Based on the robust linear mixed-effects model findings, the intercept was significant ($\beta = 3414.42$, $SE = 109.05$, $t = 31.31$), representing the mean science MSE score for the Stacking. The effects of XGBoost ($\beta = 23.42$, $SE = 2.24$, $t = 10.45$), HGB ($\beta = 25.87$, $SE = 2.24$, $t = 11.54$), LightGBM ($\beta = 22.92$, $SE = 2.24$, $t = 10.22$), and Blending ($\beta = 57.47$, $SE = 2.24$, $t = 25.63$) were all significant, indicating increases in science MSE compared to the Stacking. The random effects showed substantial variability across countries (variance = 904,255.40, $SD = 950.92$).

5 Discussion

Using machine-learning algorithms to predict student performance in ILSAs has gained significant attention in educational research. Through the utilization of machine learning techniques, it is possible to predict the performance of students and identify those who are at risk, thus enabling the implementation of timely actions to improve the academic outcomes of students (Oppong, 2023). Researchers have utilized various machine learning techniques, such as DT, NN, SVM, RF, and LR, to accurately predict students' academic achievements (Ofori et al., 2020). However, machine learning algorithms are subject to certain constraints, such as producing models characterized by higher variance, higher bias, and lower accuracy (Mienye & Sun, 2022; Mishra et al., 2022; Sun et al., 2021).

Ensemble learning models, on the other hand, enhance the outcomes of machine learning by combining multiple models, leading to better predictive accuracy in contrast to individual models (Chen et al., 2020). The main purpose of ensemble learning is to boost the performance of a model in various tasks like classification, prediction, and function approximation, while also decreasing the chances of selecting a poor model (Polikar, 2012). Bagging, boosting, and stacking represent the three primary ensemble methods extensively employed in machine learning.

In this study, the stacking method, which is an ensemble learning approach, was employed to predict the academic performance of students who participated in the 2022 administration of the PISA, one of the ILSAs. To our knowledge, this study represents the first instance of utilizing stacking to predict student performances in ILSAs. Other ensemble learning techniques, known as boosting and blending, were utilized to conduct a comparative analysis. Three boosting techniques were employed in this study: XGBoost, LightGBM, and HGB. While XGBoost and LightGBM have been commonly used boosting techniques to predict student performances in ILSAs, this study also introduced HGB, one of the boosting algorithms, and blending techniques for the first time to predict student performances in ILSAs. Employing stacking, boosting, and blending models enables the utilization of the outcome variable, student performance in ILSAs, without the requirement of dichotomization as pass or fail. Dichotomization poses challenges due to the loss of valuable information and the failure to consider within-category differences, meaning all student scores that fall either above or below the designated cut-point, which is frequently determined as the sample median, are regarded as being equal (Naggara et al., 2011). Additionally, through stacking, boosting, and blending techniques, filling in the missing values was unnecessary. Finally, we highlighted the accuracy and performance of the stacking method by comparing it with three boosting algorithms, namely XGBoost, HGB, and LightGBM, and blending.

5.1 Limitations and future research

This study is subject to several limitations. First, it relies exclusively on the PISA dataset to generate the findings. However, the stacking method can also be used to predict student performance using other ILSAs or existing various forms of educational data, including grades data, behavioral data, and demographic data, all of which provide insights into the diverse aspects of students' characteristics (Chen & Zhai, 2023). Thus, a further examination of different ILSA datasets should be conducted as part of future research. Second, our study focused on machine learning algorithms that allow different forms of boosting and blending to make the results comparable to those from the stacking method. However, it is also possible to employ more sophisticated algorithms, such as artificial neural networks and deep learning, to predict student achievement. Future studies can compare the performance of these algorithms against the stacking method. Lastly, this study used predictive modeling techniques for a regression task (i.e., predicting PISA scores on a continuous scale). However, predictive models can also be used for classification tasks where a categorical variable (e.g., at-risk students vs. others) can be predicted. Such classification models are often more challenging due to data-related issues such as class imbalance (Wongvorachan et al., 2023). Thus, future studies should also examine the performance of the stacking method in classification tasks based on ILSAs.

6 Conclusion

The present research employed three boosting algorithms, stacking, and blending to predict the academic achievement of students who took part in the 2022 administration of the PISA. Students' mathematics, reading, and science performances were predicted for all participating countries. The performances of the algorithms were evaluated using various metrics, such as MAPE, MAE, MSE, and RMSE. Among the 80 countries that participated in the study, stacking yielded the lowest metrics for most countries compared to the three boosting algorithms and blending. The robust linear mixed-effects models showed that stacking resulted in significantly lower MAPE, MAE, and MSE scores across all subjects, in contrast to the boosting and blending techniques. Therefore, when it comes to predicting student performance, stacking could be considered as a favorable option for achieving better performance, and more stable and accurate predictions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10639-024-13110-2.

Author contributions All authors contributed to the study conception, design, data analysis, writing, and editing. All authors read and approved the final manuscript.

Funding This work was not supported by any specific funding for the authors.

Data availability Data is publicly available at https://www.oecd.org/pisa/data/2022database/

Declarations

Conflict of interests The authors do not possess any competing interests that are relevant to the content discussed in this article.

References

Acıslı-Celik, S., & Yesilkanat, C. M. (2023). Predicting science achievement scores with machine learning algorithms: A case study of OECD PISA 2015–2018 data. *Neural Comput. Appl., 35*(28), 21201–21228. https://doi.org/10.1007/s00521-023-08901-6

Alam, A., & Mohanty, A. (2022). Predicting Students' Performance Employing Educational Data Mining Techniques, Machine Learning, and Learning Analytics. In: Tomar, R.S.,et al.Communication, Networks and Computing. CNC 2022. Communications in Computer and Information Science, vol 1893. Springer, Cham. https://doi.org/10.1007/978-3-031-43140-1_15

Alshareef, F., Alhakami, H., Alsubait, T., & Baz, A. (2020). Educational data mining applications and techniques.*International Journal of Advanced Computer Science and Applications,11*(4) https://doi.org/10.14569/ijacsa.2020.0110494

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

Baker, R. S. J. D. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112–118.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H. Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Educ. Inf. Technol., 28*(1), 905–971. https://doi.org/10.1007/s10639-022-11152-y

Bayirli, E. G., Kaygun, A., & Öz, E. (2023). An Analysis of PISA 2018 Mathematics Assessment for Asia-Pacific Countries Using Educational Data Mining. *Mathematics, 11*(6), 1318. https://doi.org/10.3390/math11061318

Bhutoria, A., & Aljabri, N. (2022). Patterns of cognitive returns to information and communication technology (ICT) use of 15-year-olds: Global evidence from a hierarchical linear modeling approach using PISA 2018. *Comput. Educ., 181*,. https://doi.org/10.1016/j.compedu.2022.104447

Bidegain, G., & Mujika, J. F. L. (2020). Exploring the relationship between attitudes toward science and PISA scientific performance. *Revista De Psicodidáctica (English Ed.), 25*(1), 1–12. https://doi.org/10.1016/j.psicoe.2019.08.002

Breiman, L. (1996). Stacked Regressions. *Mach. Learn., 24*, 49–64. https://doi.org/10.1007/BF00117832

Brownlee, J. (2021). *Ensemble Learning Algorithms With Python: Make Better Predictions with Bagging, Boosting, and Stacking*. San Francisco, CA, USA: Machine Learning Mastery.

Büyükkıdık, S. (2023). Influential Factors on Mathematical Literacy of Turkish Students: An Educational Data Mining Study Using PISA 2015 Data. *Psycho-Educational Research Reviews, 12*(2), 505–521. https://doi.org/10.52963/PERR_Biruni_V12.N2.10

Büyükkıdık, S., Bakırarar, B., & Bulut, O. (2018). Comparing the performance of data mining methods in classifying successful students with scientific literacy in PISA 2015. 6th International Congress on Measurement and Evaluation in Education and Psychology, Prizren, Kosova. https://doi.org/10.7939/R3KW5812Q

Cao, C., Zhang, T., & Xin, T. (2024). The effect of reading engagement on scientific literacy–an analysis based on the XGBoost method. *Front. Psychol., 15*, 1329724. https://doi.org/10.3389/fpsyg.2024.1329724

Carazo-Díaz, C., & Prieto-Valiente, L. (2024). The dramatic loss of statistical power when dichotomising continuous variables. *Rev. Neurol., 78*(1), 27–29. https://doi.org/10.33588/rn.7801.2023163

Chanamarn, N., Tamee, K., & Sittidech, P. (2016). Stacking technique for academic achievement prediction. *Int. Work. Smart Info-Media Syst. Asia (SISA 2016), no Sisa, 2016*, 14–17.

Chen, F., & Cui, Y. (2020). Investigating the relation of perceived teacher unfairness to science achievement by hierarchical linear modeling in 52 countries and economies. *Educ. Psychol., 40*(3), 273–295. https://doi.org/10.1080/01443410.2019.1652248

Chen, M., & Hastedt, D. (2022). The paradoxical relationship between students' non-cognitive factors and mathematics & science achievement using TIMSS 2015 dataset. *Stud. Educ. Eval., 73*,. https://doi.org/10.1016/j.stueduc.2022.101145

Chen, C. H., Tanaka, K., Kotera, M., & Funatsu, K. (2020). Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications. *Journal of Cheminformatics, 12*, 1–16. https://doi.org/10.1186/s13321-020-0417-9

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, Y., & Zhai, L. (2023). A comparative study on student performance prediction using machine learning.*Education and Information Technologies*, 1–19. https://doi.org/10.1007/s10639-023-11672-1

Cui, S., Yin, Y., Wang, D., Li, Z., & Wang, Y. (2021). A stacking-based ensemble learning method for earthquake casualty prediction. *Appl. Soft Comput., 101*,. https://doi.org/10.1016/j.asoc.2020.107038

D'Agostino, A., Schirripa Spagnolo, F., & Salvati, N. (2022). Studying the relationship between anxiety and school achievement: Evidence from PISA data. *Statistical Methods & Applications, 31*(1), 1–20. https://doi.org/10.1007/s10260-021-00563-9

Daza, A., Bobadilla, J., Apaza, O., & Pinto, J. (2023). Stacking ensemble learning model for predict anxiety level in university students using balancing methods. *Informatics in Medicine Unlocked, 42*,. https://doi.org/10.1016/j.imu.2023.101340

Dineva, K., & Atanasova, T. (2020). Systematic Look at Machine Learning Algorithms-Advantages, Disadvantages and Practical Applications. *International Multidisciplinary Scientific GeoConference: SGEM, 20*(2.1), 317–324. https://doi.org/10.5593/sgem2020/2.1/s07.041

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Front. Comp. Sci., 14*, 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dou, J., Yunus, A. P., Bui, D. T., Merghadi, A., Sahana, M., Zhu, Z., Chen, C.-W., Han, Z., & Pham, B. T. (2020). Improved landslide assessment using support vector machine with bagging, boosting, and stacking ensemble machine learning framework in a mountainous watershed, Japan. *Landslides, 17*, 641–658. https://doi.org/10.1007/s10346-019-01286-5

Forbes, C. T., Neumann, K., & Schiepe-Tiska, A. (2020). Patterns of inquiry-based science instruction and student science achievement in PISA 2015. *Int. J. Sci. Educ., 42*(5), 783–806. https://doi.org/10.1080/09500693.2020.1730017

Foshay, A. W., Thorndike, R. L., Hotyat, F., Pidgeon, D. A., & Walker, D. A. (1962).Educational achievements of thirteen-year olds in twelve countries: Results of an international research project, 1959–1961. Hamburg: UNESCO Institute for Education. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000131437. Access 20 Feb 2024.

Gao, S. (2014). Relationship between science teaching practices and students' achievement in Singapore, Chinese Taipei, and the US: An analysis using TIMSS 2011 data. *Front. Educ. China, 9*(4), 519–551. https://doi.org/10.3868/s110-003-014-0043-x

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In2015 international symposium on educational technology (ISET)(pp. 125–128). IEEE. https://doi.org/10.1109/ISET.2015.33

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hossain, S. M. M., & Deb, K. (2021). Plant leaf disease recognition using histogram based gradient boosting classifier. InIntelligent Computing and Optimization: Proceedings of the 3rd International Conference on Intelligent Computing and Optimization 2020 (ICO 2020)(pp. 530–545). Springer International Publishing. https://doi.org/10.1007/978-3-030-68154-8_47

Jeganathan, S., Lakshminarayanan, A. R., Ramachandran, N., & Tunze, G. B. (2022). Predicting Academic Performance of Immigrant Students Using XGBoost Regressor. *International Journal of Information Technology and Web Engineering (IJITWE), 17*(1), 1–19. https://doi.org/10.4018/IJITWE.304052

Johansson, S. (2016). International large-scale assessments: What uses, what consequences? *Educational Research, 58*(2), 139–148. https://doi.org/10.1080/00131881.2016.1165559

Kahraman, N. (2014). Cross-grade comparison of relationship between students' engagement and TIMSS 2011 Science Achievement.*Eğitim ve Bilim,39*(172). Retrieved from https://egitimvebilim.ted.org.tr/index.php/EB/article/download/2842/615. Accessed 1 Mar 2024.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances In Neural Information Processing Systems, 30. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. Accessed 8 Mar 2024.

Khine, M. S., Fraser, B. J., & Afari, E. (2020). Structural relationships between learning environments and students' non-cognitive outcomes: Secondary analysis of PISA data. *Learning Environ. Res., 23*, 395–412. https://doi.org/10.1007/s10984-020-09313-2

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the Growing Importance of International Large-Scale Assessments. In: von Davier, M., Gonzalez, E., Kirsch, I., Yamamoto, K. (eds) The Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4629-9_1

Koller, M. (2016). robustlmm: an R package for robust estimation of linear mixed-effects models. *J. Stat. Softw., 75*, 1–24. https://doi.org/10.18637/jss.v075.i06

Koyuncu, İ, & Gelbal, S. (2020). Comparison of Data Mining Classification Algorithms on Educational Data under Different Conditions. *Journal of Measurement and Evaluation in Education and Psychology, 11*(4), 325–345. https://doi.org/10.21031/epod.696664

Kunapuli, G. (2023). *Ensemble Methods for Machine Learning*. Simon and Schuster.

Lee, H., & Lee, J. W. (2021). Why East Asian students perform better in mathematics than their peers: An investigation using a machine learning approach. CAMA Working Paper No. 66/2021. https://doi.org/10.2139/ssrn.3896033

Liu, H., Chen, X., & Liu, X. (2022). Factors influencing secondary school students' reading literacy: An analysis based on XGBoost and SHAP methods. *Front. Psychol., 13*,. https://doi.org/10.3389/fpsyg.2022.948612

Manav-Demir, N., Gelgor, H. B., Oz, E., Ilhan, F., Ulucan-Altuntas, K., Tiwary, A., & Debik, E. (2024). Effluent parameters prediction of a biological nutrient removal (BNR) process using different machine learning methods: A case study. *J. Environ. Manage., 351*,. https://doi.org/10.1016/j.jenvman.2023.119899

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access, 10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Miscevic-Kadijevic, G. (2015). TIMSS 2011: Relationship between self-confidence and cognitive achievement for Serbia and Slovenia. *Revista Electrónica de Investigación Educativa, 17*(3), 109–115. https://redie.uabc.mx/redie/article/view/529/1292.

Mishra, S., Shaw, K., Mishra, D., Patil, S., Kotecha, K., Kumar, S., & Bajaj, S. (2022). Improving the accuracy of ensemble machine learning classification models using a novel bit-fusion algorithm for healthcare AI systems. *Front. Public Health, 10*,. https://doi.org/10.3389/fpubh.2022.858282

Naggara, O., Raymond, J., Guilbert, F., Roy, D., Weill, A., & Altman, D. G. (2011). Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *Am. J. Neuroradiol., 32*(3), 437–440. https://doi.org/10.3174/ajnr.A2425

Nair, P. C., Gupta, D., Devi, B. I., & Kanjirangat, V. (2023). Building an Explainable Diagnostic Classification Model for Brain Tumor using Discharge Summaries. *Procedia Computer Science, 218*, 2058–2070. https://doi.org/10.1016/j.procs.2023.01.182

Nilsen, T., & Teig, N. (2022). A systematic review of studies investigating the relationships between school climate and student outcomes in TIMSS, PISA, and PIRLS. In T. Nilsen, A. Stancel-Piątak, & J. E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education.* Springer, Cham: Springer International Handbooks of Education. https://doi.org/10.1007/978-3-030-38298-8_34-1

Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. *Journal of Information and Technology, 4*(1), 33–55. https://stratfordjournals.org/journals/index.php/Journal-of-Information-and-Techn/article/view/480.

Oppong, S. O. (2023). Predicting Students' Performance Using Machine Learning Algorithms: A Review. *Asian Journal of Research in Computer Science, 16*(3), 128–148. https://doi.org/10.9734/AJRCOS/2023/v16i3351

Örnek, F., Afari, E., & Alaam, S. A. (2023). Relationship between students' ICT interactions and science achievement in PISA 2018: The case of Türkiye.*Education and Information Technologies*, 1–23. https://doi.org/10.1007/s10639-023-12388-y

Perry, L. B., Saatcioglu, A., & Mickelson, R. A. (2022). Does school SES matter less for high-performing students than for their lower-performing peers? A quantile regression analysis of PISA 2018 Australia. *Large-Scale Assessments in Education, 10*(1), 1–29. https://doi.org/10.1186/s40536-022-00137-5

Polikar, R. (2012). Ensemble learning.*Ensemble machine learning: Methods and applications*, 1–34. https://doi.org/10.4249/scholarpedia.2776

Puah, S. (2020). Predicting Students' Academic Performance: A Comparison between Traditional MLR and Machine Learning Methods with PISA 2015. Master's Thesis, Ludwig-Maximilians-Universität München, Munich, Germany. https://doi.org/10.31234/osf.io/2yshm

R Core Team. (2023). *R: A language and environment for statistical computing [Software]*. R Foundation for Statistical Computing.

Raj, J. S., Shi, Y., Perikos, I., & Balas, V. E. (2023). Intelligent Sustainable Systems: Proceedings of ICISS 2023. https://doi.org/10.1007/978-981-99-1726-6

Rodríguez, S., Valle, A., Gironelli, L. M., Guerrero, E., Regueiro, B., & Estévez, I. (2020). Performance and well-being of native and immigrant students. Comparative analysis based on PISA 2018. *J. Adolesc., 85*, 96–105. https://doi.org/10.1016/j.adolescence.2020.10.001

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3*(1), 12–27. https://doi.org/10.1002/widm.1075

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4). https://doi.org/10.1002/widm.1249

Shi, Q. (2014). Relationship between teacher efficacy and self-reported instructional practices: An examination of five Asian countries/regions using TIMSS 2011 data. *Front. Educ. China, 9*(4), 577–602. https://doi.org/10.3868/s110-003-014-0045-x

Smyth, P., & Wolpert, D. (1997). Stacked density estimation. *Advances in neural information processing systems* (pp. 668–674)

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Comput. Educ., 157*,. https://doi.org/10.1016/j.compedu.2020.103964

Sun, Y., Li, Z., Li, X., & Zhang, J. (2021). Classifier selection and ensemble model for multi-class imbalance learning in education grants prediction. *Appl. Artif. Intell., 35*(4), 290–303. https://doi.org/10.1080/08839514.2021.1877481

Teoh, C. W., Ho, S. B., Dollmat, K. S., & Tan, C. H. (2022). Ensemble-Learning techniques for predicting student performance on video-based learning. *International Journal of Information and Education Technology, 12*(8), 741–745. https://doi.org/10.18178/ijiet.2022.12.8.1679

Torney-Purta, J., & Amadeo, J. A. (2013). International large-scale assessments: Challenges in reporting and potentials for secondary analysis. *Res. Comp. Int. Educ., 8*(3), 248–258. https://doi.org/10.2304/rcie.2013.8.3.248

Wolpert, D. H. (1992). Stacked Generalization. *Neural Netw., 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*(1), 54. https://doi.org/10.3390/info14010054

Wu, T., Zhang, W., Jiao, X., Guo, W., & Hamoud, Y. A. (2021). Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. *Comput. Electron. Agric., 184*,. https://doi.org/10.1016/j.compag.2021.106039

Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports, 4*(5). https://doi.org/10.1002/eng2.12482

Yu, C. H. (2012). Examining the relationships among academic self-concept, instrumental motivation, and TIMSS 2007 science scores: A cross-cultural comparison of five East Asian countries/regions and the United States. *Educ. Res. Eval., 18*(8), 713–731. https://doi.org/10.1080/13803611.2012.718511

Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.

Zhou, T., & Jiao, H. (2023). Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educ. Psychol. Meas., 83*(4), 831–854. https://doi.org/10.1177/00131644221117193

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ersoy Öz[1] · **Okan Bulut**[2] · **Zuhal Fatma Cellat**[1] · **Hülya Yürekli**[1]

✉ Hülya Yürekli
hyurekli@yildiz.edu.tr

[1] Department of Statistics, Yıldız Technical University, Istanbul 34220, Türkiye

[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB T6G 2G5, Canada