Neural Computing and Applications (2023) 35:21201–21228
https://doi.org/10.1007/s00521-023-08901-6

ORIGINAL ARTICLE

Predicting science achievement scores with machine learning algorithms: a case study of OECD PISA 2015–2018 data

Sibel Acıslı-Celik[1] · Cafer Mert Yesilkanat[1]

Received: 19 December 2022 / Accepted: 14 July 2023 / Published online: 3 August 2023

Abstract
In this study, the performance of machine learning methods was examined in terms of predicting the science education achievement scores of the students who took the exam for the next term, PISA 2018, and the science average scores of the countries, using PISA 2015 data. The research sample consists of a total of 67,329 students who took the PISA 2015 exam from 13 randomly selected countries (Brazil, Chinese Taipei, Dominican Republic, Estonia, Finland, Hungary, Italy, Japan, Lithuania, Luxembourg, Peru, Singapore, Türkiye). In this study, multiple linear regression, support vector regression, random forest, and extreme gradient boosting (XGBoost) machine learning algorithms were used. For the machine learning process, a randomly determined part from the PISA-2015 data of each country researched was divided as training data and the remaining part as testing data to evaluate model performance. As a result of the research, it was determined that the XGBoost algorithm showed the best performance in estimating both PISA-2015 test data and PISA-2018 science academic achievement scores in all researched countries. Furthermore, it was determined that the highest PISA-2018 science achievement scores of the students who participated in the exam, estimated by this algorithm, were in Luxembourg (r = 0.600, RMSE = 75.06, MAE = 59.97), while the lowest were in Finland (r = 0.467, RMSE = 79.38, MAE = 63.24). In addition, the average PISA-2018 science scores of the countries were estimated with the XGBoost algorithm, and the average science scores calculated for all the countries studied were estimated with very high accuracy.

Keywords Artificial intelligence · Interdisciplinary/transdisciplinary · Random forest · XGBoost

1 Introduction

Education is one of the most fundamental elements of the future of society. Developed and developing countries attach great importance to education. The Programme for International Student Assessment (PISA) is a test that tries to measure whether students are ready for real life. Furthermore, educational practices in countries that have achieved success in the examination are very important in terms of improving the quality of education and better preparing students for the future by being taken as examples by countries that have achieved little success in the examination [1]. Determining the academic success of students is extremely important in terms of both revealing the success levels of the education policies of the countries and determining the external factors affecting the specific success of these countries. For this purpose, by implementing the PISA every three years since 2000, the Organisation for Economic Cooperation and Development (OECD) has been measuring students' reading comprehension levels and ability to use maths and science skills in certain age groups in more than 90 countries [2]. In addition to the assessment of educational approaches for countries, PISA also contains important data on the development of educational processes in three-year periods. The estimation of students' academic achievement levels based on these data and the determination of the importance levels of different variables for countries are very important in terms of the development and improvement of educational processes.

It is important to determine which variables are of high importance, as the variables of individuals' characteristics have a great influence on the accuracy of the results in

✉ Cafer Mert Yesilkanat
cmyesilkanat@artvin.edu.tr

Sibel Acıslı-Celik
sacisli@artvin.edu.tr

[1] Science Teaching Department, Artvin Çoruh University, Artvin, Turkey

future forecasting and decision-making processes [3]. As necessary, educational environments and education reforms can be adjusted in order to examine these highly important variables in detail and to eliminate the differentiation between successful and less successful countries according to the results obtained. Furthermore, it is believed that the comparative analysis of the importance level of the variables of each country will make a significant contribution to the literature in terms of revealing and evaluating the dissimilar categories.

Although many studies have been conducted on PISA data in recent years, almost all of these studies have focused on due diligence for different countries and the effects of some variables on academic achievement [4–11]. However, the fact that the number of variables in PISA is high and the level of importance of the variables differ depending on the education policies of the countries reveals the difficulty of examinations with classical statistical methods. For these reasons, different and up-to-date approaches, such as machine learning, are thought to increase the interpretability skills for such large and intensive data. Furthermore, stochastic-based machine learning algorithms give more accurate and faster prediction results than traditional methods because of their ability to successfully predict non-linear relationships [12]. In this study, the prediction performance of three powerful algorithms and the results of the classical multiple linear regression (MLR) model are shown comparatively. In MLR, a single model is tried to be established in a way that creates a linear connection between the estimators and the predicted on the basis of the relationship between the variables. On the other hand, two of the machine learning algorithms examined in this study (XGBoost and Random forest) stand out as the most widely and successfully used algorithms of ensemble learning methods. In ensemble learning methods, rather than creating a single model, many models are created, and the results obtained from each model are ensured to affect the final estimation result. They do this in two different ways, such as boosting (XGBoost example) and bagging (Random Forest example) [13]. In the boosting method, the models are trained sequentially with randomly selected training sub-data from the main training data, and at the end of each training, this training sub-data is transferred back to the main training data, while the information obtained from the model training results is transferred to the next model so that the next model can make less erroneous predictions [14]. In the Bagging method, on the other hand, model training is carried out in parallel and the results obtained from each model are used in certain weights for the final estimation [15]. The third method (Support Vector Regression) used in the study offers a different machine learning approach by creating a support regression curve according to the outliers between the variables [16].

As discussed more broadly in the literature review in Sect. 1.1, the use of machine learning algorithms to predict academic achievement performance has increased considerably in recent years. Especially in PISA data, there were efforts to estimate academic success with traditional machine learning algorithms for many different countries, depending on many different variables, and significant success has been achieved in this regard compared to classical statistical methods. However, the number of studies carried out on the prediction of these success levels with machine learning algorithms with higher performance compared to their peers by examining the academic achievements of countries compared to each other is very low [17, 18]. In this context, the most important contribution of research to the literature is to predict the future PISA science achievement scores of countries by using machine learning algorithms and to reveal the importance levels of the variables in predicting the science achievement of students. In addition, the comparison of these results for countries with different proficiency levels provided a better interpretation of the research results. Furthermore, there are a limited number of studies [17–30] in the literature on predicting future science academic achievement and determining variable importance levels using existing machine learning algorithms. This situation was the main motivation for the research. Based on this perspective, the present study differs from its counterparts in terms of both the subject and the methods used, as well as revealing original results. In this respect, this study, which is one of the pioneering studies in the literature in the field, basically has five research questions (RQs) given below.

- *RQ 1.* Can the students' PISA science scores based on some specific variables be predicted with machine learning algorithms?
- *RQ 2.* Can the science academic achievement average scores of the countries in the next PISA period be predicted successfully with machine learning methods?
- *RQ 3.* To what extent does the predicted science score by machine learning algorithms differ for countries with different levels of science achievement?
- *RQ 4.* Are there significant differences between machine learning algorithms in terms of the performance of accurately predicting PISA science scores?
- *RQ 5.* What is the significance level of the variables in estimating science achievement scores and do they differ for each country?

The main purpose of this study is to estimate the average PISA 2018 science achievement scores of the students and countries participating in the exam based on the PISA 2015

exam data organized by the OECD, with machine learning algorithms. In addition, the objective was to comparatively examine the importance levels of the variables used as predictors in the study and to compare them according to countries with different success levels. This study is important in terms of predicting the science achievement scores of countries and students according to certain variables and revealing which variables will be important in the education policies to be implemented to improve these achievement scores. In this respect, it is clear that it is extremely important to compare the prediction performances of different machine learning algorithms and to determine the model with the highest performance. For this reason, the findings of this research will greatly contribute to the literature (Table 1).

1.1 Literature review

1.1.1 Machine learning for PISA

In recent years, there has been a remarkable increase in the use of machine learning methods to examine the success of students who have participated in scaled exams such as PISA [17–30]. These studies summarized in Table 2 highlight the significant variables affecting student success and the performance of various machine learning algorithms, depending on the effects of these variables. These studies have employed a wide range of machine learning algorithms, such as linear discriminant analysis (LDA), support vector machine (SVM), random forest (RF), artificial neural networks (ANN), decision trees (DT), and linear regression-based algorithms (Ridge Regression, Lasso Regression, Hierarchical Linear Modeling, Elastic Net) to predict students' PISA performance scores or identify the key factors influencing success. It is noteworthy that a significant portion of the studies given in Table 2 approach the problem as a classification type and propose solutions suitable for this feature [18–27]. In studies conducted with different country samples, although the performance of machine learning algorithms in predicting student student achievement score classes varies, it has been determined that random forest algorithm (RF), support vector machine (SVM), and artificial neural networks (ANN) consistently yield successful results in many studies [19–21, 23–25, 27]. In addition, although a single machine learning algorithm model (or two similar models) is used in some of the classification-based studies, more emphasis has been placed on identifying the factors affecting student success rather than the models' performance in predicting student success [18, 22, 26]. Alongside the classification-based studies in the literature, there are a few studies where student success scores are numerically predicted in regression-type problems. [17, 28–30]. These studies generally focused on the identification of the factors and variables affecting student success [17, 30] and determining the performance of different machine learning algorithms in predicting students' success scores [28, 29]. However, predictions of future success scores of students or countries have not been made with the models created in these few studies.

When the literature information given in this section is evaluated in general, it is seen that machine learning algorithms are successfully used comparatively with a view to classification and regression to determine the factors affecting student success and to predict students' success scores. However, these literature debates point to the need for more research into at least two critical issues. The first is the comparative determination of the performance of machine learning algorithms in predicting the future success scores of students or countries participating in the exam, and the second is the comparison of variables affecting student success for different countries according to their levels of importance. In this context, this study attempts to predict the future success scores of students or countries participating in the exam and comparatively determine the performance of machine learning algorithms.

1.1.2 Factors affecting the science score in PISA

In recent years, many studies have been conducted in which the factors affecting the science literacy scores of the students who participated in the PISA exam were examined by traditional statistical methods. The general results of some of these studies are as follows: as a result of their studies that investigated students' science literacy for Scandinavian countries according to the PISA 2000 exam results in terms of similarities and differences, Kjærnsli and Lie [31] found that some variables such as the geographical, cultural, or political context affected success. Erbaş [32] found that the factors that positively affect the science literacy of the students participating in the PISA exam from Türkiye were the teacher-student relationship, number of books in the household and participation in pre-school education, internet use, and basic computer knowledge; while Anıl [33] found that within the scope of PISA 2006, the variables most predicting the science achievement of 15-year-old students in Türkiye were respectively, the father's educational status and the attitude towards science; and Ho [34] found that according to the results of the PISA 2000–2003–2006 exams, Hong Kong students performed exceptionally well in mathematics, science, and reading, and the differences in achievement due to the socioeconomic levels of the students were less in Hong Kong compared to other countries. Anagün [35] found that the variable that affects the science literacy level of the students who took the PISA 2006 exam from Türkiye the most

Table 1 Literature summary for some studies applying machine learning to PISA data

References	Purpose of the study	Problem type	ML algorithms	Results
[17]	Analyzing the determinants of PISA-2015 test scores of students from Australia, Canada, France, Germany, Italy, Japan, Spain, United Kingdom and USA	Regression	RE-EMT	As a result of the study, it was found that the variables at the student and school level, which are significantly related to students' achievement, show significant differences between countries
[18]	To estimate the academic achievement levels of students participating in PISA 2015 from 58 countries using the support vector machine (SVM)	Classification	SVM	In this study, they found that science literacy, parents' educational status, the disciplinary environment in the classroom, and the time spent on learning variables were the predictors of success
[19]	Predicting students' PISA 2012 math performance	Classification	ANN, RF	It has been found that the ANN algorithm performs better
[20]	To estimate the variables affecting the science literacy achievement of students participating in the PISA 2015 exam from Türkiye	classification	ANN, LR	As a result of their studies, they found that the variable of time devoted to science education (SMINS) has a high importance
[21]	To investigate the relationships between German students' attitudes towards information and communication technologies (ICT) and mathematical and scientific literacy in relation to mathematical and scientific literacy in the PISA 2015 and 2018 exams	Classification	HLM, RF	As a result of the study, it has been determined that attitudes toward information and communication technologies are an important and explanatory variable for the RF model
[22]	To analyzed the variables affecting the science literacy of the students participating in the PISA 2015 exam from Türkiye	Classification	DT	As a result of the study, they concluded that the variable that most affects science literacy is the number of books in the house (ST013Q01TA)
[23]	Predicting Finnish students' math achievement	Classification	LDA, SVM, RF,	It was found that the SVM algorithm was more successful for this study
[24]	Comparative analysis of student achievement classification performance of artificial neural networks, decision trees and separation analysis methods of students participating in the PISA 2012 exam	Classification	ANN, DT	As a result of the study, it was determined that the MLPANN algorithm showed more successful results
[25]	Comparative analysis of algorithms used to predict students' PISA 2015 science literacy achievement with the help of 12 variables	Classification	DS, HT, LM, RT, RF, RLR	The most successful algorithm in the study was determined as random forest
[26]	To determine the relationship between information and communication technology (ICT) use and students' academic performance for 44 countries in PISA 2015	Classification	HLM	The study found that national ICT skills, students' ICT availability at school, and student interest in ICT use had a more positive impact on student academic performance
[27]	Using Multilayer Perceptron Neural Networks and Random Forest methods, to determine the factors affecting PISA 2015 mathematical literacy and to compare the prediction performance of both methods	Classification	RF, ANN	They found that the random forest (RF) algorithm was more successful
[28]	To estimate the factors affecting the mathematical literacy of students from Singapore, Japan, Norway, America, Türkiye and the Dominican Republic who took the PISA-2015 exam	Regression	DT, ANN	As a result of the study, it was determined that the variables that affect the mathematical literacy of countries with different proficiency levels differ
[29]	Predicting students' science score performance in PISA-2015	Regression	RR, LR, ENET, RT, MLR	As a result of the study, it was found that all models used in the study had very similar performance in predicting students' science performance in PISA

Table 1 (continued)

References	Purpose of the study	Problem type	ML algorithms	Results
[30]	To identify the main factors affecting the academic performance of schools in Tunisia that took the PISA 2012 exam	Regression	RT, RF	The most successful algorithm was determined to be the RF in the study

In alphabetical order, *ANN* artificial neural network, *DS* decision stump, *DT* decision trees, *ENET* elastic net, *HLM* hierarchical linear modelling, *HT* Hoeffding-Tree, *LDA* Linear Discriminant Analysis, *LM* Logistic Model, *LR* Lasso Regression, *MLR* Multiple Linear Regression, *RE-EMT* Mixed-Effects Regression Tree, *RF* Random Forest, *RLR* Ridge logistic Regression, *RR* Ridge Regression, *RT* Random Tree, *SVM* Support Vector Machines

Table 2 The Science literacy scores of the countries investigated in the study according to PISA 2015 data, and the number of students participating in PISA (Note-1: Students with missing variables are not included in this number. Note-2: Countries are listed alphabetically according to Qualification levels.)

Countries	PISA-2015		
	Qualification levels	PV2SCIE mean ± SD	Student numbers
Chinese Taipai	Level-3	532 ± 95	6726
Estonia	Level-3	534 ± 85	4781
Finland	Level-3	531 ± 91	4955
Japan	Level-3	538 ± 90	4985
Singapore	Level-3	556 ± 97	4727
Hungary	Level-2	477 ± 89	3866
Italy	Level-2	481 ± 84	9192
Lithuania	Level-2	475 ± 89	5133
Luxembourg	Level-2	483 ± 95	3577
Türkiye	Level-2	425 ± 76	4679
Brazil	Level-1a	401 ± 88	8341
Peru	Level-1a	397 ± 73	4482
Dominican Republic	Level-1b	332 ± 74	1885
Total			67,329

PV2SCIE PISA 2015 science achievement score, *SD* standart deviation

is "time spent for learning", while their attitudes towards learning science did not affect their science literacy. As a result of their studies using HLM and Mplus software, Acar and Öğretmen [36] observed that the performance of Turkish students in the 2006-PISA science test differs according to student and school levels. Using the example of PISA 2006 Hong Kong, Sun et al. [37] investigated the factors affecting the science achievement of 15-year-old students. At the end of the study, it was determined that male students, students from high socioeconomic status (SES) families, students with higher motivation and higher self-efficacy, and students whose parents value science highly are more likely to succeed in science. Sälzer and Heine [38] investigated whether there is a relationship between student achievement and absenteeism according to the PISA 2012 results, while Kahraman and Çelik [39] found that attending kindergarten, high educational status of parents, and a high number of computers and books in a household positively affected students' success in science literacy. Chen and Cui [40], through hierarchical linear modeling analysis, found that perceived teacher injustice was negatively related to science achievement as a result of their study that investigated the gender, student-school, and the economic, social, and cultural statuses of the most influential variables in the science achievement of students from 52 countries who took the PISA 2015 exam. As a result of their study in which they compared the scientific achievements of students who took the PISA 2015 exam from Singapore and Türkiye, Karakoç-Alatli [41] found that interest in science, science self-efficacy, and research-based science practices in both countries had a positive effect on success. Rohatgi and Scherer [42] examined the science achievement of students participating in PISA 2015 from Norway, and as a result of the study, the highest correlation was found between TEACHUP and PERFEED variables. Antonelli et al. [43] found that EMOSUPS and

HOMEPOS variables were predictors when they examined mathematics, reading and science scores for students participating in PISA 2015 from Brazil. Atasoy et al. [44] investigated the variables affecting the science achievement of students in Türkiye, the USA and South Korea, using PISA 2018 data. As a result of their studies, it was determined that the variables of gender and socioeconomic status had a positive effect on science achievement. Karakus et al. [45] examined the mathematics, science and reading achievement performance of first and second generation immigrant students using PISA 2018 data. According to the findings of their study, they determined that the variables of parental support (EMOSUPS), teacher enthusiasm and the adaptation of instruction were associated with an increase in academic performance.

1.2 Purpose of the research

The main purpose of this study is to use the data obtained from the PISA 2015 exam, which was organized by the OECD and focused on science literacy in 2015, to estimate the PISA 2018 science literacy levels and average country scores of the students participating in the research with multiple linear regression (MLR), support vector regression (SVR), random forest (RF), and extreme gradient boosting (XGBoost) machine learning algorithms. For this purpose, variables that affect students' PISA 2015 science literacy levels for 13 randomly selected countries (Brazil (BRA), Chinese Taipai (TAP), Dominican Republic (DOM), Estonia (EST), Finland (FIN), Hungary (HUN), Italy (ITA), Japan (JPN), Lithuania (LTU), Luxembourg (LUX), Peru (PER), Singapore (SGP), Türkiye (TUR)) from different proficiency levels were examined separately and predictions were made at certain optimization levels with each machine learning algorithm evaluated in the study. In addition, 18 variables that are thought to directly affect science achievement were used as estimators of the PISA 2018 science literacy score for each country examined in the study, and whether these variables have a significant effect on students' science literacy achievements and the importance levels of these variables were determined separately for each country. By identifying the important factors affecting the achievement of science literacy with machine learning algorithms, the study also aims to contribute to the development of studies on the subject. In addition to these, in this study, SVR, RF, and XGBoost machine learning algorithms, which is popular in recent years, were compared in terms of their prediction performances, as well as the classical statistical method MLR. It's a pioneering study on predicting the future science education achievement scores with current and high-performance machine learning algorithms. Considering all these aspects, it explicitly differs from similar studies.

2 Materials and methods

2.1 The data and software resources

OECD's Programme for International Student Assessment, widely known as the PISA test, measures 15-year-old students' reading comprehension levels and their ability to use their math and science skills. This evaluation exam, which is held every three years, was created for the first time in 1997 and its first application was made in 2000. In the PISA application, each semester focuses on a specific course, but students are also tested from other core courses. In this context, the focus was on the science course in 2015.

All PISA-2015 and PISA-2018 data used in the study have been obtained from the OECD's own data repository which is publicly available [46, 47]. The population of the research consists of 519,334 students from 72 countries who took the PISA 2015 exam. From this sample, 13 countries (67,329 students) were randomly selected. According to the results of the PISA-2015 science achievement evaluation, the qualification levels of the 5 of these countries (Chinese Taipei, Estonia, Finland, Japan, and Singapore) are Level 3, while 5 of them (Hungary, Italy, Lithuania, Luxembourg, and Türkiye) are Level 2, and 3 of them (Brazil, Peru, and the Dominican Republic) are Level 1. Table 2 shows the average PISA-2015 science literacy scores of the countries included in the sample, their proficiency levels, and the number of students participating from these countries. In this context, the study is limited to 13 different countries and 18 variables examined in the study.

In this study, the Training/Test split, which is one of the reliable validation methods in the literature, was used [48]. PISA-2015 data of each selected country was split into two parts as training and testing data. This means that each country's training and testing data are evaluated differently from each other. 75% of the randomly determined part of the whole data set for the countries was used as training data for the training of machine learning models, and the remaining 25% was split as testing data to test the learning performances of the model with validation. These subgroups of data were randomly determined to represent all science scores of countries at an appropriate level. Since there are quite a lot of sampling numbers in the PISA-2015 data, the missing data, which were detected in a small number, were directly excluded from the calculation in the data preprocessing phase.

The complete study is conducted using the R statistical programming environment [49, 50] and while kernlab is used for the SVR algorithm [51]; for the RF algorithm, randomForest [52]; for the XGBoost algorithm, xgboost [53] for data preparation, separation and optimization caret

[54], and for cross validation, scatter diagrams and data visualization openair [55] ve ggplot2 [56] R library packages were used.

2.2 The model process steps and performance metrics

In order to systematically analyze the research questions of the study mentioned in Sect. 1, a 7-stage process was carried out in this study. These stages can be summarized as follows;

1. Thirteen countries that have participated in the PISA 2015 test were randomly selected.
2. Variables that are thought to affect science achievement were determined in accordance with the literature in Sect. 1.1. Accordingly, 18 variables were determined as predictors in the study. These variables are listed in Table 3. Also as Science score, PV2SCIE coded score levels were used. Further, a detailed description of each variable is provided in supplementary material-1.
3. PISA-2015 data of each selected country was split into two parts as training and testing data. 75% of the randomly determined part of the whole data set for the countries was used as training data for the training of machine learning models, and the remaining 25% was split as testing data to test the learning performances of the model with validation.
4. For the learning models, besides the traditional MLR calculation; SVR, RF, and XGBoost algorithms, which perform the learning process and are explained in detail in Sect. 2.3, are used. Through these four methods, the supervised learning process for the training data set has been carried out, and hyperparameter variables have been optimized with the grid-search method for the greatest prediction performance (tenfold cross-validation and Number of resampling iterations is 10). In the grid-search method, all possible combination results of the possible values of each hyperparameter are listed. Then the machine learning model is reconstructed according to each combination, thus determining the hyperparameter coefficients with the best prediction performance [57]. Detailed performance outputs determined for each model in the optimization process are given in supplementary material-2. At the end of the learning process, the testing data that was not introduced to the model was estimated with each model and these results were compared with the actual data.
5. To compare the predicted science scores using machine learning models with the actual results and to determine the performance levels, mean absolute error (MAE, Eq. 1), root mean square error (*RMSE*, Eq. 2), and Pearson's correlation coefficient (r, Eq. 3) performance metrics were used. These performance metrics were calculated as follows:

Table 3 Variables used in the research on science literacy

PISA-codes	Explanation	References
ST011Q04TA	In your home: A computer you can use for school work	[25, 26, 29, 33, 36, 40, 42, 43]
ST004Q01TA	Gender	[17, 22, 27–29, 36–38, 40, 44, 45]
ST013Q01TA	Total number of books found at home	[19, 22, 25, 26, 28, 29, 36, 39, 42, 43]
ST123Q02NA	Parental support for educational efforts and achievements	[22, 25, 43, 45]
MISCED	Mother's education (ISCED)	[18, 22, 25, 28, 29, 33, 39, 42]
FISCED	Father's education (ISCED) Mother's education (ISCED)	[18, 22, 25, 28, 29, 33, 39, 42]
HISCED	Highest education of parents (ISCED)	[25, 28, 29, 33, 37, 39, 42]
ST097Q03TA	Class discipline in science classes	[18, 20, 25, 28, 41, 42]
SMINS	Learning time—Science (minutes per week)	[18, 20, 25, 28, 35, 36]
TMINS	Learning time—in total (minutes per week)	[18, 20, 25, 28, 35, 36]
BELONG	Subjective well-being: Sense of Belonging to School	[25, 27, 29, 35, 42, 43]
EMOSUPS	Parents emotional support	[22, 25, 27, 37, 42, 43, 45]
HOMEPOS	Home possessions	[22, 25, 26, 28, 29, 39, 40, 42, 43]
WEALTH	Family wealth	[25, 26, 29, 32]
ESCS	Index of economic, social and cultural status	[25–29, 31, 33, 36, 37, 40, 42, 44, 45]
ADINST	Teacher adaptation	[18, 32, 41]
PERFEED	Received Feedback	[18, 25, 27, 42]
TEACHSUP	Teacher support in a science classes of students' choice	[18, 25, 41, 42]

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|A_i - P_i| \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(A_i - P_i)^2} \tag{2}$$

$$r = \frac{\sum_{i=1}^{n}(A_i - \bar{A})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{n}(A_i - \bar{A})}\sqrt{\sum_{i=1}^{n}(P_i - \bar{P})}} \tag{3}$$

where n is the total number of data, A_i and P_i denotes actual data and predicted value of i-th sample, where $\overline{A}$ and $\overline{P}$ are the mean of the actual data and predicted values, respectively. For the model with high predictive performance, *RMSE* and *MAE* metrics should be close to zero [58]. Also Pearson's correlation coefficient can be categorized as weak correlations ($r < 0.39$), moderate correlations ($0.40 < r < 0.69$), strong correlations ($0.70 < r < 89$) and very high correlations ($r \geq 0.90$) [59, 60].

6. After the model with the best prediction performance level was determined, the future PISA-2018 Science achievement scores of the countries where the education process was carried out with the PISA-2015 data were estimated and compared with the real achievement scores with validation diagrams. In addition, the distribution of the real science scores of the countries and the distribution of the science scores estimated by the model were investigated.
7. The importance levels of the predictor variables according to the machine learning model, which offered the best prediction performance, were determined for each country examined in the research.

2.3 Machine learning models

In this study, the levels of science achievement of PISA 2018 for 13 countries randomly selected among the countries participating in the PISA-2015 exam were estimated using the conventional approach of multiple linear regression (MLR), as well as machine learning algorithms of support vector regression (SVR), random forest (RF), and extreme gradient boosting (XGBoost). These models, except for XGBoost, are often used in similar studies, as described in Sect. 1.1.1. The main reason for choosing these algorithms is that they are both compatible with the structure of the problem and can determine a model-specific variable importance metric, especially with RF and XGBoost [61].

2.3.1 Support vector regression (SVR)

The support vector machine (SVM) [62] algorithm, which is successfully used in both classification and regression problems is named support vector regression (SVR) when the algorithm is used for regression problems [63]. Unlike linear regression, model errors in SVR are tried to be kept within a certain threshold. Thus, the negative effects of the outlier data on the model are reduced. In the SVR algorithm trained with the structural risk minimization principle, the input variables are mapped into a higher dimensional space using the kernel function [64, 65]. Accordingly, the SVR function can be defined as follows.

$$f(x) = \omega\varphi(x) + b + \varepsilon \tag{4}$$

where $\varphi(x)$ is the kernel function used for non-linear mapping; x is the input vector for the model; ω and b are the regulation parameters of the function and ε indicate the level of upward and downward deviation from the SVR model separation curve. The main purpose of the SVR algorithm is to determine $f(x)$ the optimized parameters of the function ($\varphi(x)$ and b) so that the ε deviation stays within the range. These parameters can be found by solving the minimizing problem shown in Eq. 5 according to the limits in Eq. 6.

$$\mathrm{Min}\left[\frac{1}{2}\omega^2 + C\sum_{i=1}^{m}\xi_i + \xi_i^{\star}\right] \tag{5}$$

$$\text{Subject to}\begin{cases} y_i - \omega\varphi(x_i) - b \leq \varepsilon + \xi \\ \omega\varphi(x_i) + b - y_i \leq \varepsilon + \xi_i^{\star} \\ \xi_i, \xi_i^{\star} \leq 0, 1, 2, 3, \cdots, m \end{cases} \tag{6}$$

where ξ and $\xi^{\star}$ are slack variables which are indicators of outlier values above and below ε-tube and C is the penalty coefficient which specifies the threshold range for the SVR model. A detailed description of the SVR algorithm can be found in the literature [65–67].

2.3.2 Random forest (RF)

Originally developed by Brieman [68], the RF algorithm is defined as a combination of bagging [69] and random subspace [70] approaches [71]. RF is a black-box modeling technique based on the principle of random creation of multiple decision trees and generating random forests from these decision trees that meet certain criteria [72, 73]. In the RF algorithm, firstly, 2/3 of the whole data set is randomly split as training data. Multiple decision trees are created with bootstrap samples [71, 74] Ultimately, the final estimate of RF is the average of all results obtained from each individual tree. Since the training is carried out with bootstrap samples, the RF algorithm is resistant to overfitting [75]. Also, this model can be used to accurately

describe complex nonlinear relationships due to the high randomness effect. [76].

2.3.3 Extreme gradient boosting (XGBoost)

First proposed by Chen and Guestrin [53], XGboost algorithm is a scaleable, optimized and faster processing implementation of the gradient boosting machine (GBM) [77] framework. XGBoost differs from GBM by processes such as parallel computing, tree and leaf pruning, regularisation, and optimization [78].

The objective function of XGBoost consists of the sum of $l(\cdot)$ (loss function) and $\Omega(\cdot)$ (regulation term) [79] as shown in Eq. 7.

$$L = l(\cdot) + \Omega(\cdot) = \sum_{i=0}^{N} l(y_i, \widehat{y}_i) + \sum_{k=0}^{K} \Omega(f_k) \tag{7}$$

where $\widehat{y}_i$ is the estimated value for the ith sample; y_i is the ith actual value; N is the number of samples; K is the number of decision trees and f_k is the model of the k th tree. The regulation term that prevents the overfitting level of the model from rising $\Omega(\cdot)$ can be defined as shown in Eq. 8.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \|\omega_j\|^2 \tag{8}$$

where T is the number of leaves in a decision tree; ω_j is the score on the j th leaf; γ and λ are defined as penalty terms.

Iteration process is applied to minimize the objective function (L) given in Eq. 7 and to determine the best model parameters. Each new function series created by this process is created by optimizing the errors of the previous function. Hence, the new objective function in the tth iteration will be as shown in Eq. 9.

$$L^{(t)} = \sum_{i=0}^{N} l\left(y_i, \widehat{y}_i^{(t-1)} + f_t(x_i)\right) + \gamma T \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{9}$$

As a result, a second-order Taylor expansion is applied to Eq. 9, so that the objective function can be optimized to develop a high-performance prediction model [79–81].

3 Results and discussion

3.1 Pisa-2015 data: science performance score

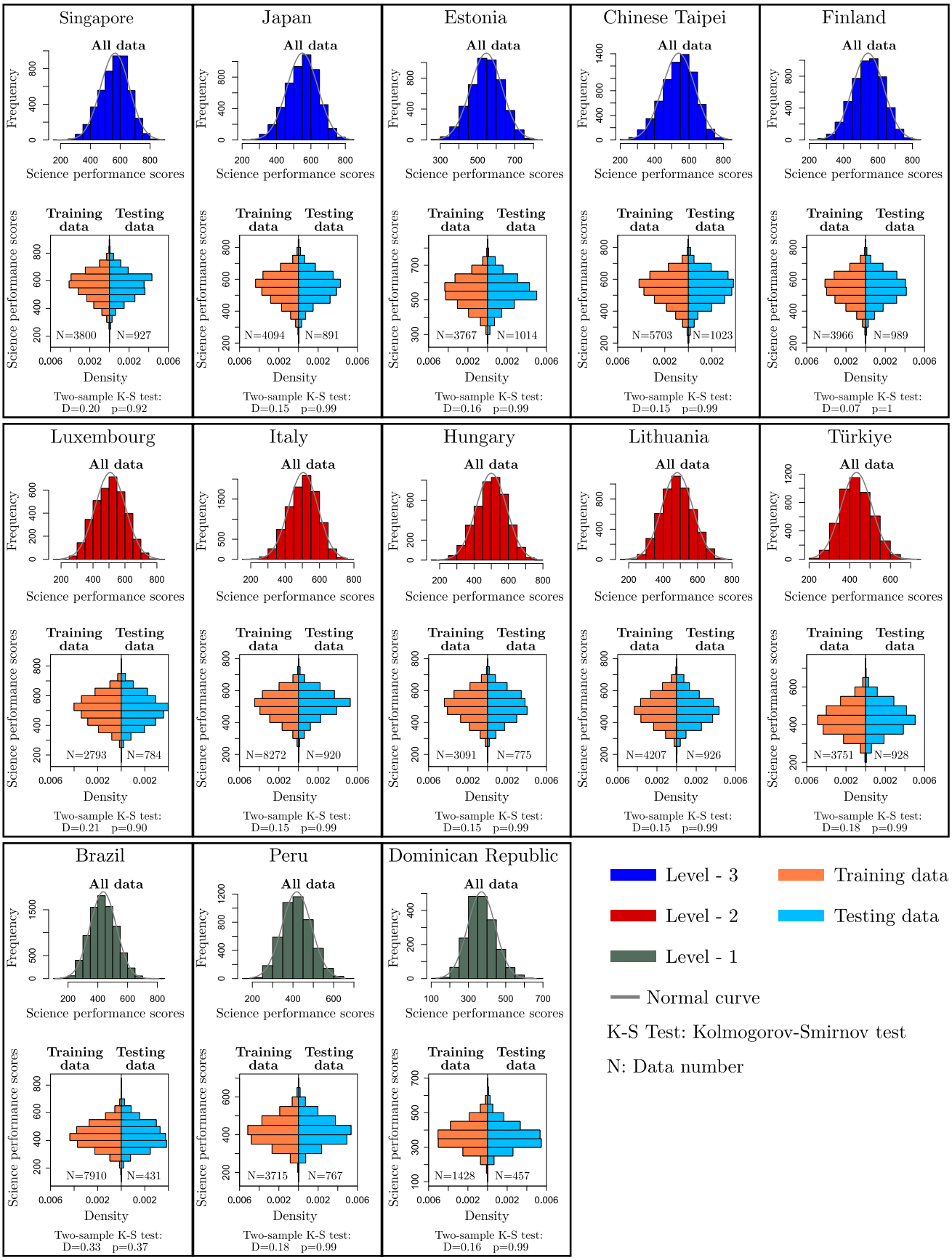
In Fig. 1, general distributions of Science performance scores of the students that participated in PISA 2015 in countries randomly selected for this study are shown with histogram diagrams for all data, training data, and testing data. When Fig. 1 is examined for the general histograms created with all data, it's seen that the science scores of all 13 countries were normally distributed. This indicates that the probability distribution of the science score, which is a random variable, will follow a normal distribution. Also, when these results are evaluated together with Table 2, Singapore has the highest mean science score (556 ± 97), while the Dominican Republic has the lowest (332 ± 74). Additionally, standard deviation levels, which are an indicator of the level of distribution according to the mean of science performance scores, were determined at close levels for all countries examined, except Türkiye, Peru, and the Dominican Republic.

When Fig. 1 is examined for the histograms created with training and testing data, it can be seen that distributions similar to the histograms created with all data are obtained. Furthermore, according to the results of the two-sample Kolmogorov–Smirnov test, it has been determined that the training data and the testing data for all countries investigated in the study have the same data distribution structure and therefore their level of representation is high ($p > 0.05$ for all countries).

3.2 Machine learning models: training and testing

In Table 4, the performance metrics of the models obtained depending on the hyperparameter values that give the highest performance scores at the end of the optimization process detailed in SM2 are given for both training data and testing data. Accordingly, it was determined that the XGBoost algorithm showed the best prediction performance in all the countries studied. Although SVR and RF algorithms performed very close to each other, they were below XGBoost in terms of prediction performance. It was determined that the prediction performance of the MLR classical statistical method is quite low compared to other popular machine learning methods. The results obtained from Table 4 are summarized in a simplified way in Fig. 2. In this figure, the changes in Pearson's r correlation coefficients obtained for each model used in the study for both training data and testing data by country are given comparatively. Figure 2 clearly shows that the XGBoost algorithm has a higher correlation coefficient for each country than the other models. The correlation coefficient is a measure of how closely two variables are linearly related. For this reason, it has been determined that there is a high correlation between the prediction results obtained from the XGBoost algorithm and the actual data.

When Table 4 and Fig. 2 are evaluated together, the highest performance metric score according to the XGBoost algorithm is in Luxembourg for both training data and testing data (for training data $r = 0.644$, RMSE = 72.46, MAE = 58.31; for testing data $r = 0.678$,


Singapore
All data
Frequency
Science performance scores
Training data
Testing data
N=3800
N=927
Density
Two-sample K-S test:
D=0.20 p=0.92
Japan
All data
Frequency
Science performance scores
Training data
Testing data
N=4094
N=891
Density
Two-sample K-S test:
D=0.15 p=0.99
Estonia
All data
Frequency
Science performance scores
Training data
Testing data
N=3767
N=1014
Density
Two-sample K-S test:
D=0.16 p=0.99
Chinese Taipei
All data
Frequency
Science performance scores
Training data
Testing data
N=5703
N=1023
Density
Two-sample K-S test:
D=0.15 p=0.99
Finland
All data
Frequency
Science performance scores
Training data
Testing data
N=3966
N=989
Density
Two-sample K-S test:
D=0.07 p=1
Luxembourg
All data
Frequency
Science performance scores
Training data
Testing data
N=2793
N=784
Density
Two-sample K-S test:
D=0.21 p=0.90
Italy
All data
Frequency
Science performance scores
Training data
Testing data
N=8272
N=920
Density
Two-sample K-S test:
D=0.15 p=0.99
Hungary
All data
Frequency
Science performance scores
Training data
Testing data
N=3091
N=775
Density
Two-sample K-S test:
D=0.15 p=0.99
Lithuania
All data
Frequency
Science performance scores
Training data
Testing data
N=4207
N=926
Density
Two-sample K-S test:
D=0.15 p=0.99
Türkiye
All data
Frequency
Science performance scores
Training data
Testing data
N=3751
N=928
Density
Two-sample K-S test:
D=0.18 p=0.99
Brazil
All data
Frequency
Science performance scores
Training data
Testing data
N=7910
N=431
Density
Two-sample K-S test:
D=0.33 p=0.37
Peru
All data
Frequency
Science performance scores
Training data
Testing data
N=3715
N=767
Density
Two-sample K-S test:
D=0.18 p=0.99
Dominican Republic
All data
Frequency
Science performance scores
Training data
Testing data
N=1428
N=457
Density
Two-sample K-S test:
D=0.16 p=0.99
Level - 3
Level - 2
Level - 1
Training data
Testing data
Normal curve
K-S Test: Kolmogorov-Smirnov test
N: Data number

◀**Fig. 1** All data, training data, and testing data histograms of PISA-2015 science achievement scores of 13 countries investigated in the study

RMSE = 70.83, MAE = 57.06), the lowest performance metric score was in Finland (for training data $r = 0.498$, RMSE = 78.79, MAE = 63.08) and for testing data in Estonia ($r = 0.442$, RMSE = 76.37, MAE = 60.44). According to these results, the level of correlation between the predicted values and the actual data was determined as moderate correlations ($0.40 < r < 0.69$).

In Fig. 3, validation diagrams can be seen in the hexagonal density structure created according to the XGBoost algorithm, which has shown the best prediction performance. Validation diagrams determined with all other models are given in supplementary material-3, comparatively. In Fig. 3 diagrams, with the XGBoost algorithm, whose learning process was performed and optimized, each data in the testing data were estimated separately and the prediction results were compared with the actual data. Since the number of data to be estimated in this study is large, and therefore it is thought that it can better characterize the relationship between the estimated values and the actual values, a hexagonal density diagram was preferred. When Fig. 3 is examined for testing data, it is seen that the scores of the students whose scores are close to the real science achievement average scores of the countries are estimated at a very successful rate. In addition, the predicted science achievement scores of a significant part of the students who got higher and lower than the average science score for all countries surveyed were generally determined at the levels of distribution density range. However, although the numbers are small, predictive performance decreases (dark blue color) at extreme values (minimum and maximum) of science scores in countries (especially Finland, Hungary, Japan and Italy). Since this situation increased the range of error, it partially increased the RMSE values for the model. It is thought that the main reason for this high amount of error observed in the extreme data is different levels of importance of the predictive variables used for machine learning training. However, the high accuracy in the prediction results of the scores of the students who have scores close to the average values is remarkable (dark red color). In these areas, lower RMSE values have been obtained. As a result, according to the findings obtained from Fig. 3 (PISA-2015), the XGBoost algorithm has produced appropriate and acceptable results in estimating the science scores of a great majority of the students who took the exam in 13 different countries, and the science score averages of the countries for the year they took the exam.

3.3 A different approach: Validation

In this part of the research, the training process was carried out with a different approach to compare multiple algorithms and further verify the descriptive statistics. This approach, unlike the one described in Sect. 3.2, involves considering the training process and all data together, without separating countries. For this, 25% (16,832 students) of the total data used in the study (67,329 students) were randomly chosen and simplified. This process was repeated three times and three different sample data sets, named sample-1, sample-2, and sample-3, were created. Afterward, these simplified data sets were randomly divided into two main sub-data groups, 75% training data (12,624 students) and 25% testing data (4208 students). The same four machine learning algorithms (MLR, SVR, RF, and XGBoost) were used in the training process and the success of the models was evaluated using statistical metrics according to their performance in predicting the testing data. Figure 4 shows the validation diagrams obtained as a result of estimating the actual PISA score (testing) data of each sample data set with machine learning algorithms. Accordingly, the XGBoost algorithm showed the highest performance for each sample, similar to the previous results. The performance metrics of the XGBoost algorithm were calculated as RMSE = 79.7, MAE = 62.9 and $r = 0.624$ for sample-1, RMSE = 79.5, MAE = 63.2 and $r = 0.615$ for sample-2, and RMSE = 78.1, MAE = 62.7 and $r = 0.640$ for sample-3. In addition, performance metrics close to each other were found for all of the sample classes in other algorithms (Fig. 4). When the findings are evaluated according to the countries, the estimation result performances of some outliers are insufficient, especially in the 1st and 3rd qualification levels countries, but the data close to the average values of the countries have been estimated with high accuracy. This result is seen in more detail in Fig. 5, which shows the error scatter diagram of the prediction scores obtained by the XGBoost algorithm for each sample group according to the qualification levels of the countries. According to this figure, the error distributions for each sample group and each qualification levels vary in similar intervals, and the biggest error levels were determined at the 1st and 3rd qualification levels. The main reason for this is thought to be due to the high number of outliers deviating from the mean value in these qualification level groups. In addition, the majority of the data with low error for all three proficiency levels of all sample groups remained within the probability contour at the 95% confidence level. As a result, it has been determined that the estimated PISA scores given by the XGBoost machine learning model created are quite successful in representing the real values.

Table 4 Performance indicators of optimized models for training and testing data (PISA-2015)

Country	Models	Training data			Testing data			Hyperparameters
		r^*	RMSE	MAE	r^*	RMSE	MAE	
BRAZIL	MLR	0.546	73.42	58.75	0.591	75.46	61.47	
(BRA)	SVR	0.574	71.80	57.27	0.630	72.66	58.95	σ = 0.037; ε = 0.1; C = 0.25
A = 7910	RF	0.581	71.46	57.07	0.627	73.19	59.17	mtry = 6; ntree = 400
B = 431	**XGBoost**	**0.581**	**71.40**	**57.01**	**0.645**	**71.47**	**57.68**	nrou = 30; α = 0.2; λ = 50
CHINESE TAIPAI	MLR	0.555	79.18	63.12	0.515	81.87	65.35	
(TAP)	SVR	0.573	78.13	62.15	0.545	80.14	63.99	σ = 0.037; ε = 0.1; C = 0.25
A = 5703	RF	0.581	77.58	61.89	0.554	79.55	63.09	mtry = 5; ntree = 500
B = 1023	**XGBoost**	**0.604**	**75.96**	**60.38**	**0.582**	**77.75**	**61.88**	nrou = 100; α = 40; λ = 1500
DOMINICAN REPUBLIC (DOM)	MLR	0.488	63.72	50.69	0.468	66.29	51.23	
A = 1428	SVR	0.535	62.12	49.29	0.513	64.40	50.46	σ = 0.037; ε = 0.1; C = 0.25
B = 457	RF	0.532	61.95	49.11	0.527	63.88	49.54	mtry = 2; ntree = 800
	XGBoost	**0.561**	**61.25**	**48.69**	**0.536**	**63.53**	**49.42**	nrou = 100; α = 10; λ = 500
ESTONIA	MLR	0.502	73.23	58.50	0.432	78.12	61.37	
(EST)	SVR	0.511	72.82	58.32	0.470	76.55	60.69	σ = 0.037; ε = 0.1; C = 0.25
A = 3767	RF	0.514	72.83	58.33	0.458	76.98	60.75	mtry = 5; ntree = 500
B = 1014	**XGBoost**	**0.515**	**72.79**	**58.07**	**0.472**	**76.37**	**60.44**	nrou = 20; α = 0.2; λ = 60
FINLAND	MLR	0.488	79.25	63.39	0.535	76.57	61.11	
(FIN)	SVR	0.493	79.20	63.42	0.525	77.13	61.51	σ = 0.040; ε = 0.1; C = 0.25
A = 3966	RF	0.492	79.49	63.75	0.522	77.73	62.00	mtry = 2; ntree = 600
B = 989	**XGBoost**	**0.498**	**78.79**	**63.09**	**0.538**	**75.97**	**59.90**	nrou = 20; α = 20; λ = 50
HUNGARY	MLR	0.612	69.71	55.63	0.610	71.73	58.30	
(HUN)	SVR	0.619	69.34	55.20	0.624	70.75	57.12	σ = 0.039; ε = 0.1; C = 0.5
A = 3091	RF	0.620	69.70	55.70	0.627	70.56	57.16	mtry = 3; ntree = 700
B = 775	**XGBoost**	**0.632**	**68.44**	**54.26**	**0.643**	**70.27**	**57.03**	nrou = 100; α = 100; λ = 1000
ITALY	MLR	0.503	73.26	58.69	0.525	72.29	58.11	
(ITA)	SVR	0.537	71.61	57.11	0.541	71.23	57.17	σ = 0.039; ε = 0.1; C = 0.25
A = 8272	RF	0.539	71.72	57.60	0.545	71.45	57.45	mtry = 4; ntree = 700
B = 920	**XGBoost**	**0.545**	**71.17**	**56.70**	**0.549**	**71.03**	**56.81**	nrou = 50; = α1000; λ = 200
JAPAN	MLR	0.493	78.26	62.34	0.542	75.87	59.58	
(JPN)	SVR	0.503	77.89	62.14	0.560	74.87	58.90	σ = 0.037;ε = 0.1; C = 0.25
A = 4094	RF	0.514	77.54	61.96	0.560	75.36	59.45	mtry = 4; ntree = 1200
B = 891	**XGBoost**	**0.522**	**76.79**	**61.09**	**0.573**	**73.94**	**57.71**	nrou = 50; = α15; λ = 500
LITHUANIA	MLR	0.497	76.83	61.91	0.483	79.19	63.49	
(LTU)	SVR	0.517	75.81	61.24	0.519	77.33	61.79	σ = 0.039; ε = 0.1; C = 0.25
A = 4207	RF	0.524	75.66	61.19	0.506	78.15	62.19	mtry = 4; ntree = 1000
B = 926	**XGBoost**	**0.534**	**75.00**	**60.37**	**0.532**	**76.67**	**60.94**	nrou = 50; = α3000; λ = 20
LUXEMBOURG	MLR	0.616	74.40	59.92	0.641	73.76	58.65	
(LUX)	SVR	0.628	73.82	59.28	0.664	71.87	58.34	σ = 0.042; ε = 0.1; C = 0.25
A = 2793	RF	0.630	73.60	59.09	0.664	72.35	58.49	mtry = 5; ntree = 750
B = 784	**XGBoost**	**0.644**	**72.46**	**58.31**	**0.678**	**70.83**	**57.06**	nrou = 50; α = 100; λ = 300
PERU	MLR	0.531	61.53	49.31	0.567	59.29	47.32	
(PER)	SVR	0.548	60.89	48.60	0.581	58.58	46.19	σ = 0.038; ε = 0.1; C = 0.5
A = 3715	RF	0.554	60.59	48.35	0.580	58.69	46.51	mtry = 4; ntree = 500
B = 767	**XGBoost**	**0.563**	**60.17**	**47.97**	**0.584**	**57.92**	**46.10**	nrou = 50; = α1000; λ = 300

Table 4 (continued)

Country	Models	Training data			Testing data			Hyperparameters
		r*	RMSE	MAE	r*	RMSE	MAE	
SINGAPORE	MLR	0.608	76.91	60.83	0.578	78.83	62.42	
(SGP)	SVR	0.625	75.83	60.08	0.607	76.75	60.99	σ = 0.040; ε = 0.1; C = 0.25
A = 3800	RF	0.624	75.87	60.05	0.623	75.54	60.12	mtry = 9; ntree = 500
B = 927	**XGBoost**	**0.633**	**75.26**	**59.24**	**0.625**	**75.11**	**59.68**	nrou = 35; α = 1500; λ = 20
TÜRKİYE	MLR	0.524	64.78	52.09	0.506	66.82	53.35	–
(TUR)	SVR	0.528	64.63	51.73	0.528	65.84	52.98	σ = 0.038; ε = 0.1; C = 0.25
A = 3751	RF	0.541	64.27	51.49	0.530	65.74	52.65	mtry = 3; ntree = 500
B = 928	**XGBoost**	**0.558**	**63.28**	**50.42**	**0.544**	**65.13**	**52.07**	nrou = 30; α = 0.5; λ= 50

*$p < 0.001$

The bold font indicates the best-performing model

A is training data number; B is testing data number

ntree = number of trees, mtry = Splite number on the node, C = The penalty coefficient, σ = Distribution parameter for Gaussian radial basis, ε = the level of upward and downward deviation of the SVR model separation curve, nrou = Max number of boosting iterations, α = L1 regularization term on weights, λ = L2 regularization term on weights

The results given in Fig. 5 are shown with supplementary material-4 separately for each country investigated in the study.

3.4 PISA-2018 data: estimating

In this part of the research, after determining the machine learning process with training data based on PISA-2015 data and the performance scores of this learning level with testing data, the process of estimating PISA-2018 science scores was carried out. The PISA 2018 Science scores of the randomly determined countries in PISA-2015 were estimated with the help of machine learning algorithms using the same variables. The estimation performance indicators for each machine learning model are summarized in Table 5. Accordingly, results similar to the learning process created with the PISA-2015 data were obtained in the PISA 2018 estimation results. According to Table 5, it has been determined that the XGBoost algorithm is quite successful compared to other methods. Furthermore, it has been determined that the MLR approach, which is a classical statistical method, has the lowest vertical prediction performance for each country compared to other machine learning algorithms. In addition, the estimation results of the country average score of science achievement for the researched countries by machine learning algorithms are shown in Table 6. This Table, similarly to the results obtained in Table 5, shows that the averages of the real science scores are estimated with higher accuracy with the XGBoost algorithm. Moreover, it was estimated that the science score estimates calculated with the XGBoost algorithm would increase for the countries at level 1 according to the science score they received in 2015, but the increase in these countries was higher than the predicted level. In addition to these, the level of change in the mean science score for all countries surveyed (partial decrease in the mean science score of the Chinese Taipai and increase in the mean science score of all other countries) was determined with high accuracy. This is an important piece of evidence showing that the future prediction performance of the machine learning model is accurate and high. In addition, the increase in average science achievement scores for countries at level 1 was more than the expected (predicted) level. For this reason, while the average science scores of countries with Qualification levels 3 and 2 are estimated with very high accuracy, There was a decrease in the accuracy of the estimation results for those at level 1.

In Fig. 6, results obtained from Table 5 are summarized in a simplified way. In this figure, Pearson's r correlation coefficient values of the prediction performances obtained with each model are shown comparatively for the countries. Accordingly, it is explicitly seen that the XGBoost algorithm's ability to predict PISA-2018 science achievement scores is higher for each country than other models.

When Table 5 and Fig. 6 are evaluated together, the success level of the predictive values of PISA-2018 science achievement scores according to the XGBoost algorithm is highest in Luxembourg (r = 0.600, RMSE = 75.06, MAE = 59.97), and lowest in Finland (r = 0.467, RMSE = 79.38, MAE = 63.24). According to these results, it was determined that the correlation level between the predictive values obtained for the PISA-2018 science score and the actual data was at the level of moderate correlations ($0.40 < r < 0.69$).

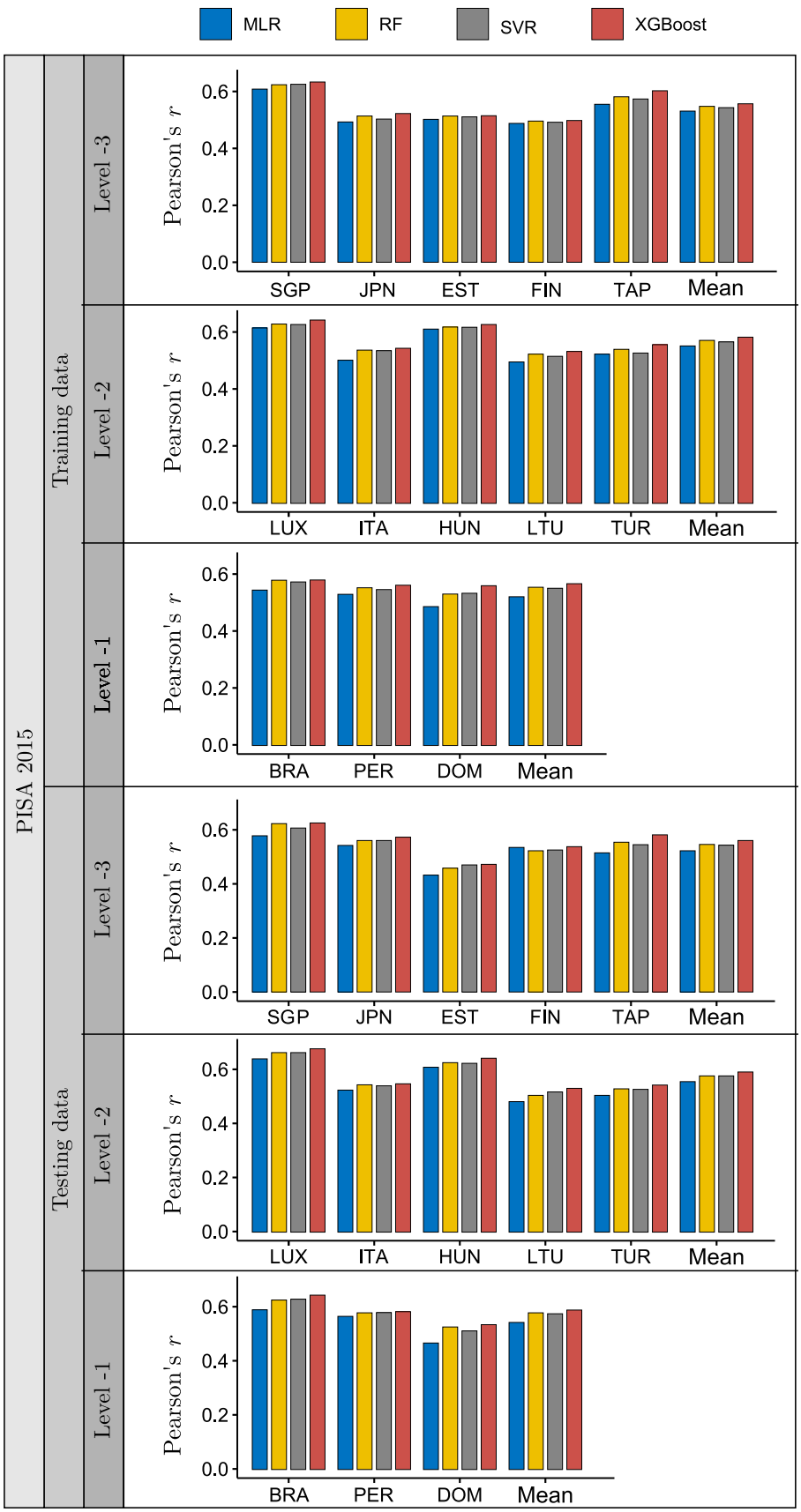


Fig. 2 Comparative evaluation of the prediction performances of machine learning models for each country

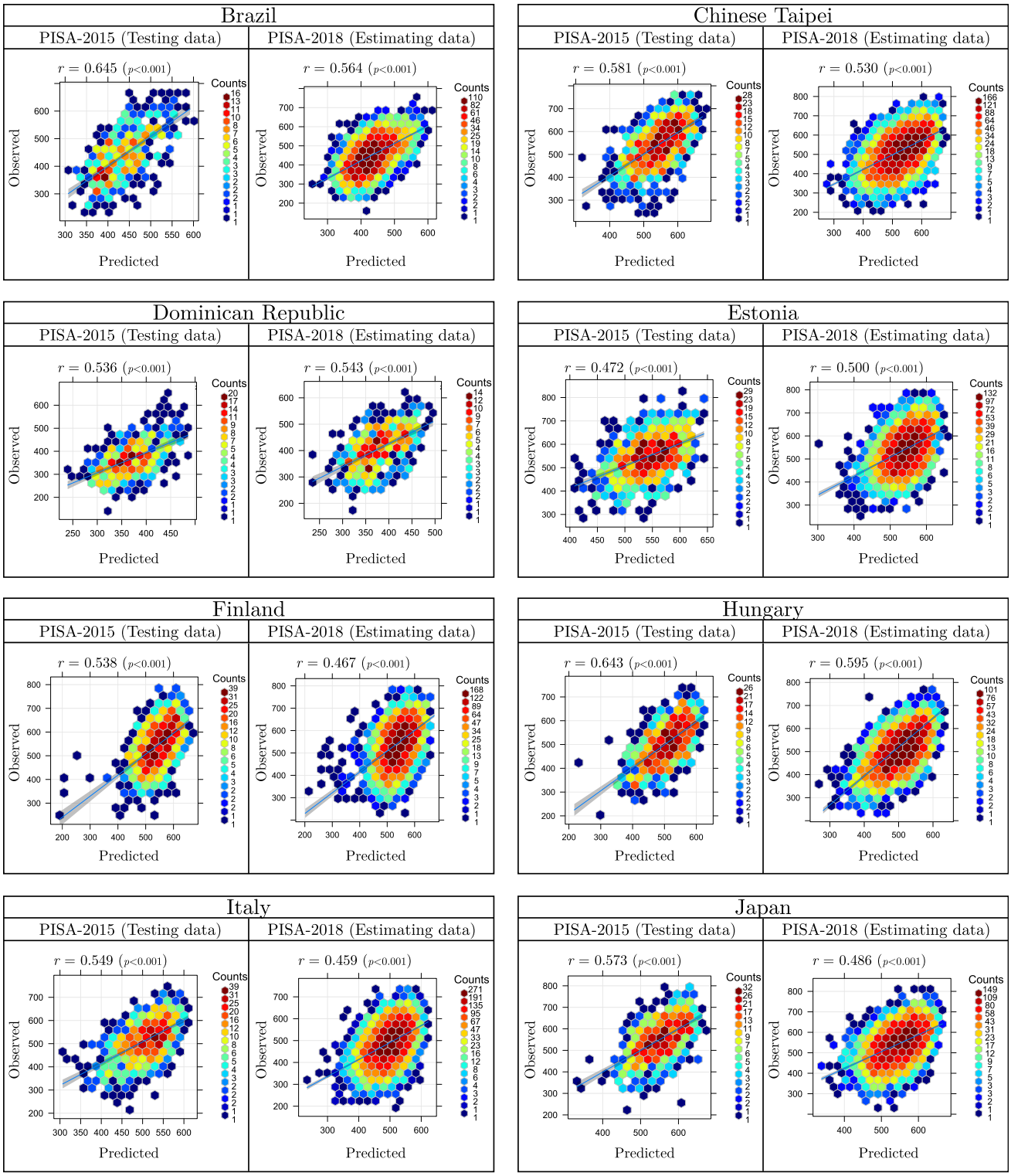


Fig. 3 Comparison of Science performance scores estimated by the XGBoost machine learning algorithm with actual results for testing data and estimating data

When Fig. 3 is examined for the estimating data (PISA - 2018) again, it is seen that the relationship between the estimated Science score scores of the students who took the exam and the actual score is at an acceptable level, except for a few outliers. According to the hexagonal density diagram of the PISA-2018 data for all countries studied,

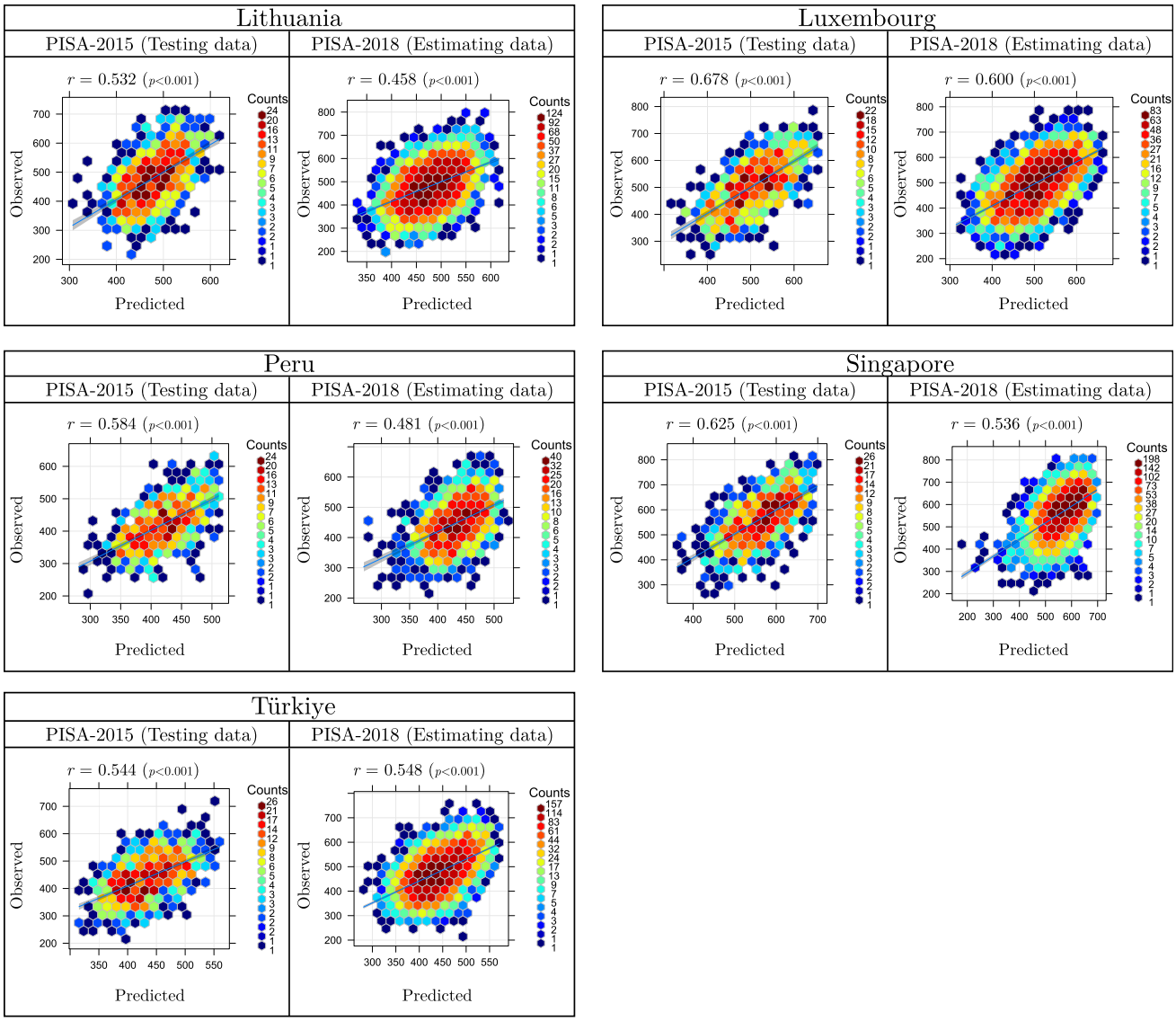


Fig. 3 continued

estimation results of the science scores of students who came close to the average level were found highly accurate, similar to the results obtained from the Testing data (dark red color). Again, similar to the results obtained from the testing data, it is seen that the prediction performance decreases, and the estimation error increases at the minimum and maximum levels of science scores (dark blue color). However, according to the number of Counts, which is the indicator of the hexagonal density diagram, it was determined that these high erroneous results were very few. These results show that with the machine learning model, the individual success of a large number of students who score close to the average scores of the countries can be estimated very close to the real result. However, the same machine learning model could not accurately predict the individual success of a small number of students who scored very different from the average scores of countries. From this point of view, it is clear that the model results should be evaluated according to the average scores of the countries, and the study should be interpreted according to these limitations.

In Fig. 7, the histograms of the science scores estimated by the XGBoost machine learning algorithm are given in comparison to the histogram distributions of PISA-2018 actual science scores of all researched countries. According to the two-sample Kolmogorov–Smirnov test, there was no significant difference between the distributions of the estimated PISA-2018 science achievement scores for all countries and the actual results ($p > 0.05$). In other words, it was determined that the science scores estimated by the machine learning model showed a significantly similar distribution with the real values in all countries studied.

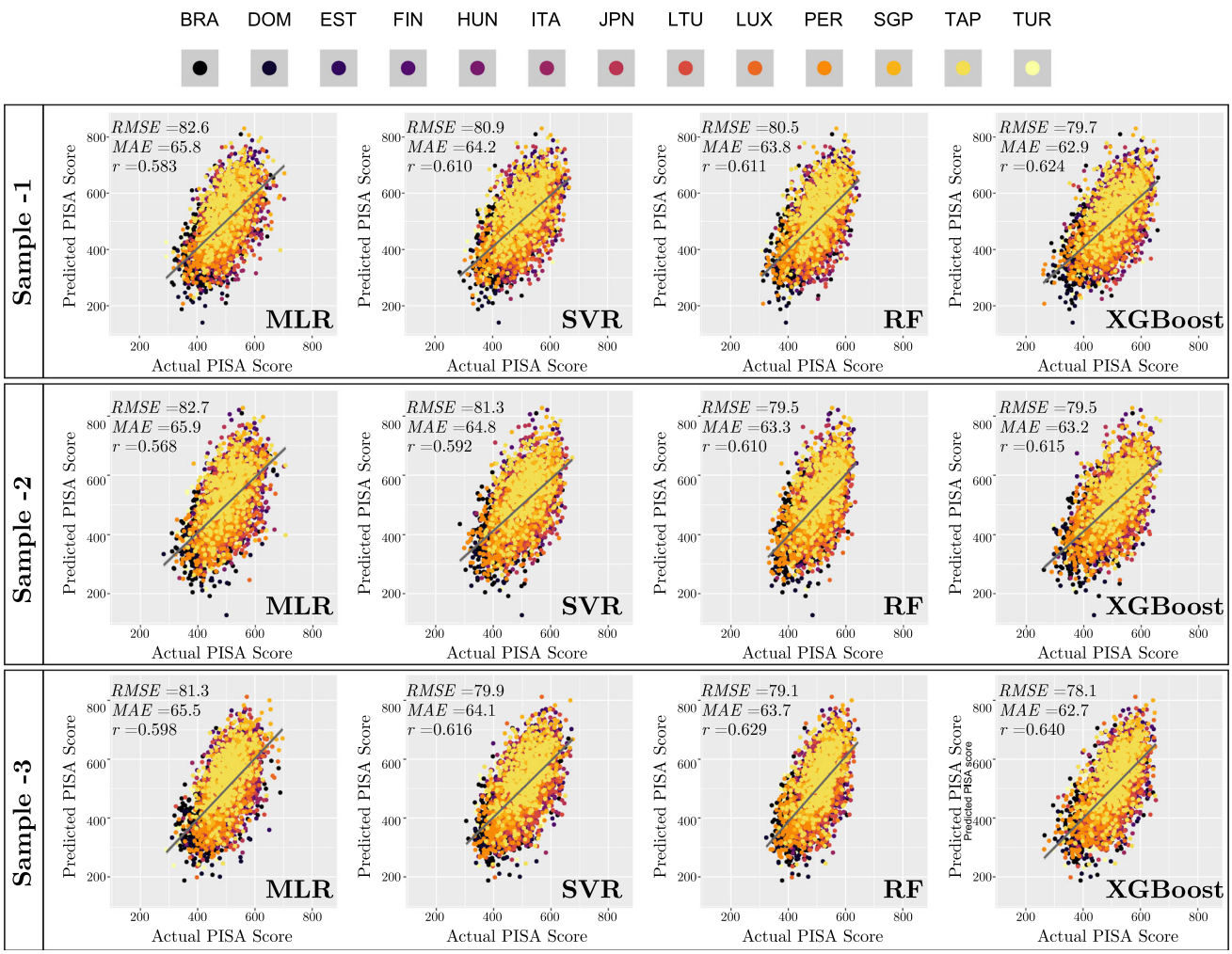


Fig. 4 Prediction performances and cross-validation diagrams for machine learning models built according to all data

Also, in Fig. 7, it is seen that the performance success of the XGBoost estimation results does not differ according to the qualification levels of the countries. As a result, according to all the findings obtained, it was determined that the XGBoost algorithm produced highly accurate and acceptable results in predicting the science scores of the students who took the exam in the researched countries and the science score averages of the countries for the future.

3.5 Performances of models: Evaluation

Whether there is a significant difference between the performance average scores of the models used in the study was tested with Tukey multiple comparisons. In this test, the correlation coefficients obtained as the result of each model were compared. According to the results given in Table 7 (testing data), there is a significant difference ($p < 0.05$) between the correlation results obtained from XGBoost, which gives the highest performance value, and MLR, which shows the lowest performance, for both data sets (training and testing data) ($p < 0.05$). In other words, the prediction values obtained with the XGBoost algorithm were determined to be significantly more successful than the prediction values obtained with the MLR. Also, whether there is a significant difference in the prediction performance of the XGBoost algorithm among the qualification levels was examined with the two-sample Tukey test, and it was determined that there is no significant difference among qualification levels of the performance success obtained from this machine learning algorithm ($p > 0.05$). Results of this test are shown in Table 7, (testing data).

Tukey multiple comparisons were used to test whether there was a significant difference between the algorithms examined in the study in terms of predicting PISA-2018 science achievement scores. With the help of this test, the correlation coefficients obtained as the result of each model were compared. According to the results given in Table 7 (estimating data), a significant difference was found between the correlation results obtained from XGBoost,

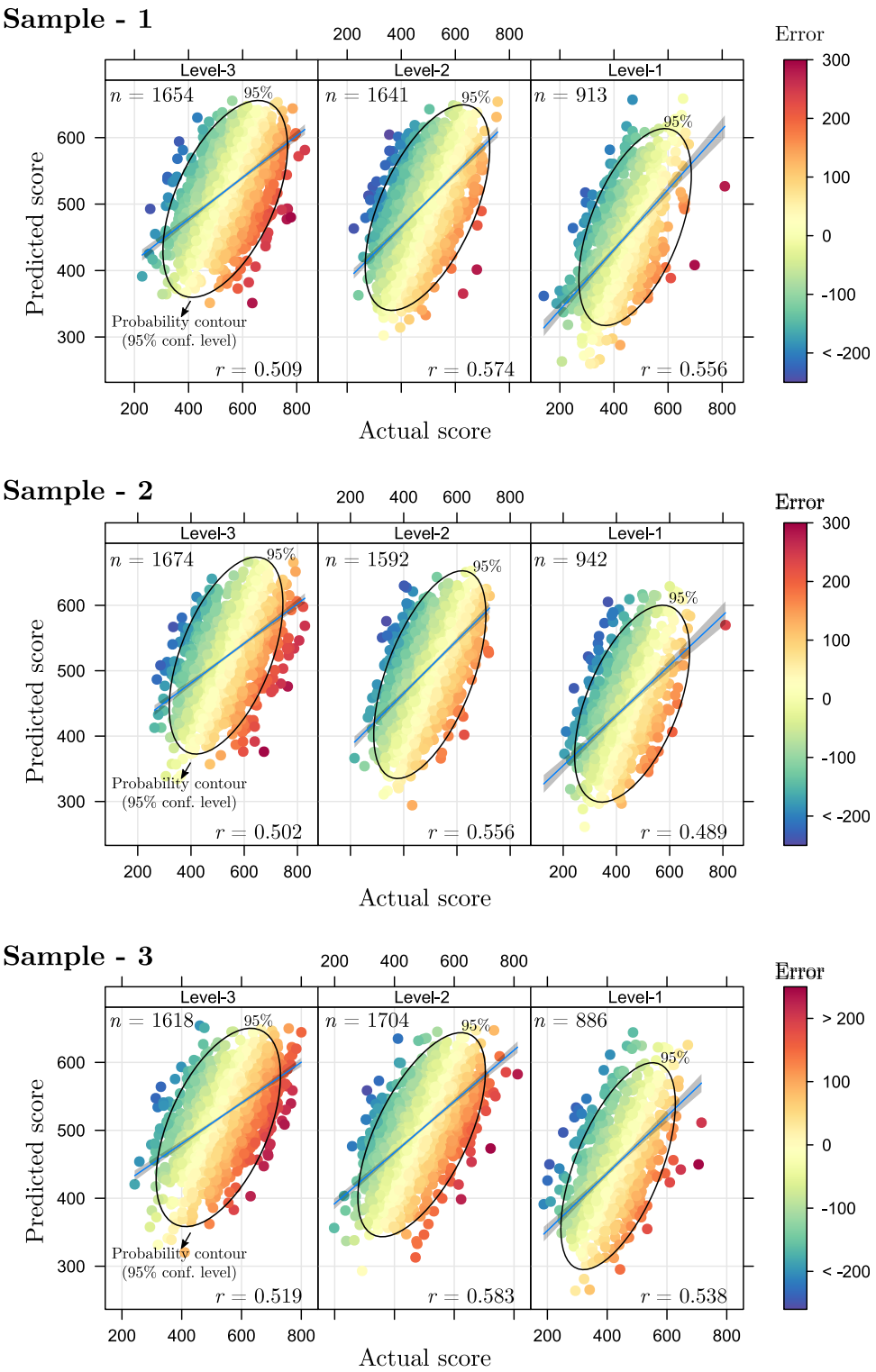


Fig. 5 Prediction performances and error scatter diagrams according to proficiency levels for the XGBoost model built according to all data

which has the highest performance in predicting the PISA-2018 science achievement score, and the correlation results obtained from MLR, which has the lowest performance ($p < 0.05$). According to this result, which was obtained similar to the PISA-2015 training and testing data, it was determined that the estimation values created with the XGBoost algorithm were significantly more successful than the estimation values obtained with the MLR. Similarly, whether there is a significant difference in the prediction performance of the XGBoost algorithm between the qualifications levels was investigated using the Tukey test with two samples and it was determined that there was no

Table 5 Estimation performance of the machine learning models for the PISA—2018 data

Country	Models	Performance indicators		
		r^*	RMSE	MAE
BRAZIL	MLR	0.546	75.46	60.45
(3485)	SVR	0.552	75.39	60.22
	RF	0.560	74.40	59.82
	XGBoost	0.564	**74.09**	**59.46**
CHINESE TAIPAI	MLR	0.492	82.95	65.82
(6296)	SVR	0.515	81.68	64.80
	RF	0.525	81.24	64.34
	XGBoost	0.530	**79.96**	**63.22**
DOMINICAN REPUBLIC	MLR	0.493	73.12	59.40
(452)	SVR	0.539	71.69	58.58
	RF	0.524	71.59	58.97
	XGBoost	0.543	**71.41**	**58.59**
ESTONIA	MLR	0.444	76.51	61.39
(4108)	SVR	0.455	76.02	60.94
	RF	0.456	75.77	60.66
	XGBoost	0.500	**64.85**	**53.31**
FINLAND	MLR	0.451	79.81	63.75
(4373)	SVR	0.448	79.95	63.66
	RF	0.453	79.50	63.40
	XGBoost	0.467	**79.38**	**63.24**
HUNGARY	MLR	0.551	73.20	58.90
(3630)	SVR	0.584	71.27	57.25
	RF	0.589	71.00	56.93
	XGBoost	0.595	**70.67**	**57.03**
ITALY	MLR	0.397	76.46	60.83
(7481)	SVR	0.444	74.72	59.52
	RF	0.445	74.67	59.27
	XGBoost	0.459	**73.77**	**58.30**
JAPAN	MLR	0.447	78.56	62.72
(4833)	SVR	0.477	77.36	61.83
	RF	0.485	76.71	61.37
	XGBoost	0.486	**76.69**	**61.17**
LITHUANIA	MLR	0.416	81.31	64.77
(4866)	SVR	0.439	80.28	64.35
	RF	0.442	79.97	64.31
	XGBoost	0.458	**79.03**	**64.19**
LUXEMBOURG	MLR	0.567	77.25	61.86
(3416)	SVR	0.581	76.61	61.26
	RF	0.591	75.65	60.56
	XGBoost	0.600	**75.06**	**59.97**
PERU	MLR	0.438	71.12	57.23
(1369)	SVR	0.479	69.83	55.80
	RF	0.465	69.70	55.72
	XGBoost	0.481	**69.08**	**55.42**
SINGAPORE	MLR	0.477	82.42	65.05
(5004)	SVR	0.515	79.79	63.12
	RF	0.525	78.96	62.04
	XGBoost	0.536	**78.12**	**61.87**
TÜRKİYE	MLR	0.516	79.84	64.40
(5550)	SVR	0.539	79.30	63.88
	RF	0.539	78.33	62.93
	XGBoost	0.548	**77.90**	**62.56**

$^*p < 0.001$, Note: The number of students participating in PISA-2018 in countries is given in parentheses

The bold font indicates the best-performing model

significant difference between the prediction performance obtained from the XGBoost algorithm and the qualifications levels ($p > 0.05$). Results of this test are shown in Table 7 (estimating data).

3.6 Effect of variables: importance level

In order to determine the importance levels of the variables in tree-based models, the gain of prediction model must be determined. Gain is the amount of accuracy that a feature or variable brings to the model. In tree models, every gain of every feature of every tree is taken into account, and then the gain per variable for the final model is allocated. As a result, it is determined that the variable with the larger amount of gain added to the model is at a more important level for generating new predictions. Thus, the contribution of each variable to the model accuracy can be ranked as a percentage according to their importance level.

Figure 8 shows the percentile importance levels of the variables for the XGBoost algorithm, which shows the best prediction results and performance. Accordingly, as can be seen in Fig. 8 (blue bars), SMINS, ESCS, HOMEPOS and TMINS variables are at the top for Singapore sample, which is one of the countries in Level 3; FISCED, MISCED, HISCED and ST123Q02NA variables were found to be of low importance; For the Japan sample, while SMNS, ESCS, TMINS and WEALTH variables are at the top; FISCED, ST011Q04TA, HISCED and ST123Q02NA variables were of low importance; While ST013Q01TA, TMINS, ESCS and PERFEED variables are at the top for the Estonian sample; ST123Q02NA, HISCED, MISCED and ST011Q04TA variables were at low significance level; While ST013Q01TA, SMINS, TMINS and ESCS variables are at the top for the Chinese Taipei sample; While ST004Q01TA, FISCED, ST011Q04TA and MISCED variables are at low importance level and for Finland

Table 6 According to PISA 2018 data, the actual values of the science literacy scores of the countries examined in the study and their average estimates according to machine learning algorithms. (Note: Missing data were excluded from the estimations and were not included in the calculation)

Countries	Number of person	qualification levels	Observed	Predicted			
			PV2SCIE (mean ± sd)	MLR (mean ± sd)	SVR (mean ± sd)	RF (mean ± sd)	XGBoost (mean ± sd)
Chinese Taipai	6296	3. Level	520 ± 94	531 ± 57	527 ± 58	524 ± 56	522 ± 65
Estonia	4108	3. Level	542 ± 85	548 ± 45	546 ± 47	544 ± 47	543 ± 51
Finland	4373	3. Level	536 ± 89	533 ± 44	538 ± 44	540 ± 34	536 ± 49
Japan	4883	3. Level	542 ± 87	540 ± 43	541 ± 44	541 ± 39	542 ± 47
Singapore	5004	3. Level	566 ± 92	560 ± 59	563 ± 58	566 ± 57	566 ± 65
Hungary	3630	2. Level	503 ± 87	500 ± 52	498 ± 53	502 ± 46	503 ± 56
Italy	7481	2. Level	489 ± 83	497 ± 41	495 ± 45	499 ± 38	492 ± 48
Lithuania	4866	2. Level	483 ± 89	487 ± 44	480 ± 44	482 ± 39	483 ± 47
Luxembourg	3416	2. Level	501 ± 93	505 ± 58	506 ± 57	504 ± 52	501 ± 62
Türkiye	5550	2. Level	468 ± 81	439 ± 42	439 ± 44	440 ± 37	454 ± 48
Brazil	3485	1a. Level	451 ± 88	433 ± 49	435 ± 51	435 ± 48	442 ± 53
Peru	1369	1a. Level	446 ± 77	430 ± 37	432 ± 39	434 ± 35	438 ± 40
Dominican Republic	452	1b. Level	408 ± 77	375 ± 37	380 ± 41	376 ± 37	385 ± 47

PV2SCIE PISA 2018 science achievement score

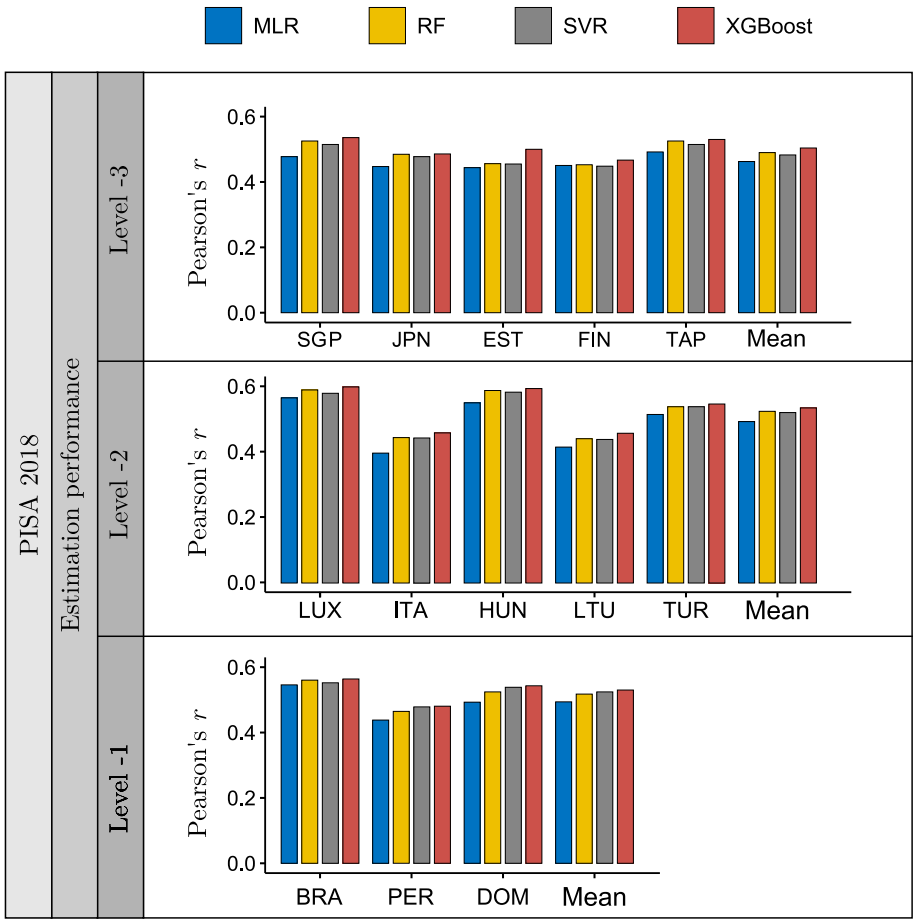


Fig. 6 Comparative evaluation of the estimation performance of machine learning models for PISA-2018 science scores of the researched countries

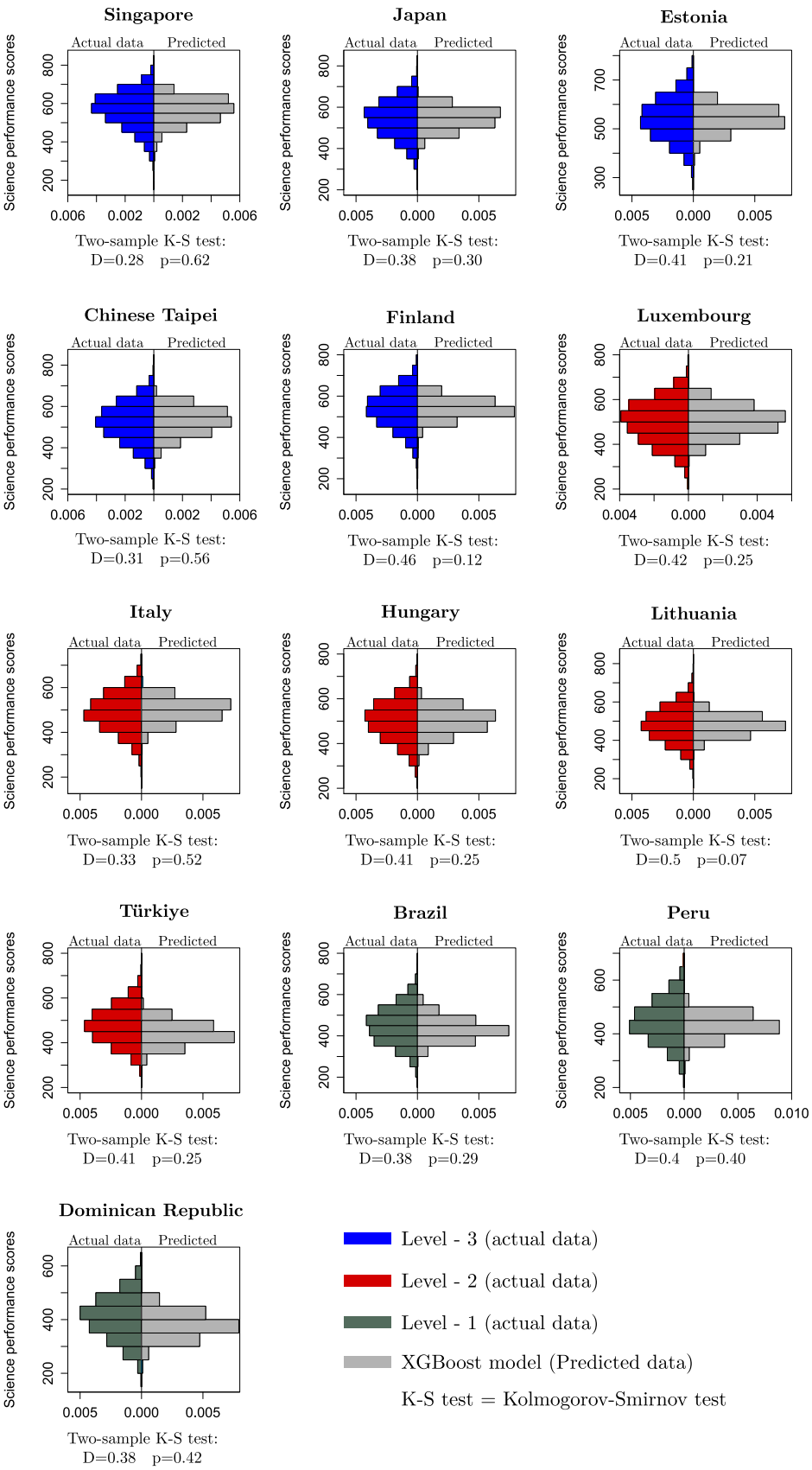


Fig. 7 Comparison of the histograms and distributions of countries' actual and predicted PISA-2018 science score values

sample, ST013Q01TA, SMINS, ESCS and PERFEED variables are at the top; HISCED, ST004Q01TA, ST123Q02NA and ST011Q04TA variables were determined to be at low significance level. Similarly, as can be seen in Fig. 8 (red bars), ST013Q01TA, ESCS, TMINS and SMNS variables are at the top for the sample of Luxembourg, which is one of the Level 2 countries; FISCED, HISCED, ST011Q04TA and ST123Q02NA variables were of low importance; For the Italy sample, the variables ST013Q01TA, TMINS, PERFEED and SMNS are at the top; MISCED, FISCED, ST011Q04TA and ST123Q02NA variables were of low importance; While ST013Q01TA, ESCS, TMINS and PERFEED variables are at the top for Hungary sample; ST123Q02NA, HISCED, FISCED and ST011Q04TA variables were at low significance level; For Lithuania sample, ESCS, ST013Q01TA, TEACHSUP and TMINS variables are at the top; FISCED, ST004Q01TA, MISCED and HISCED variables were found to be of low importance; For Türkiye sample, ST013Q01TA, WEALTH, TMINS and ESCS variables are at the top; It was determined that ST004Q01TA, FISCED, ST123Q02NA and HISCED variables were at low significance level. Finally, as can be seen in Fig. 8 (grey bars), SMINS, ESCS, TMINS and HOMEPOS variables are at the top for the Brazil sample, which is one of the Level 1 countries; FISCED, ST004Q01TA, ST011Q04TA and ST123Q02NA variables were of low importance; For the Peru sample, ESCS, BELONG, TMINS and PERFEED variables are at the top; ST011Q04TA, ST004Q01TA, EMOSUPS and ST123Q02NA variables were found to be of low importance; For the Dominican Republic sample, TMINS, HOMEPOS, BELONG and ESCS variables are at the top; HISCED, ST123Q02NA, ST004Q01TA and ST011Q04TA variables were found to be of low importance.

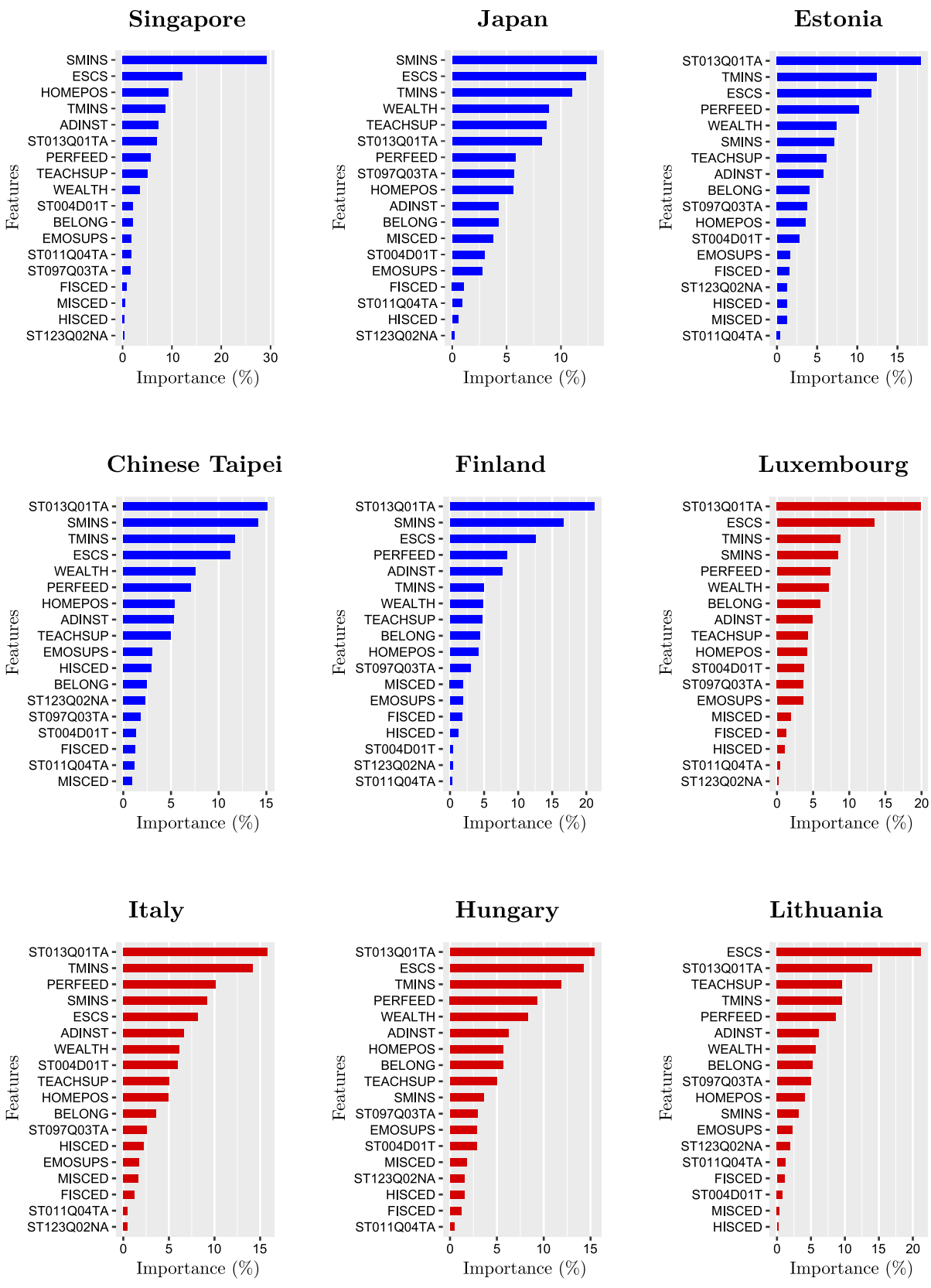
Fig. 8 Importance levels of the variables (as percentile) for the XGBoost algorithm, which shows the best prediction results and performance

As a result, SMINS, TMINS, ESCS variables rank high in terms of importance in all 13 countries participating in the study, while FISCED, MISCED, HISCED and ST123Q02NA variables were determined at low importance. According to this, although the variables of time spent learning science per week, total learning time and socio-economic level are highly important variables in terms of predicting the science achievement of countries for the future, it is observed that the variables of father education level, mother education level, highest education level of parents and parental support were found to be of low importance. In addition to these results, no significant differences were detected between the importance levels of the variables in terms of qualification levels. However, it is a remarkable result that the ST013Q01TA variable was of high importance in the 2nd and 3rd level countries, but remained at low and medium level in the 1st level countries. Another interesting result is the HOMEPOS variable. Although this variable was generally of medium importance, it was found to be of high importance in the Dominican Republic (12.4%), Singapore (9.8%), and Brazil (8.78%). Furthermore, the WEALTH variable was

Table 7 Tukey multiple comparisons results for models and qualifications levels

	Comparison of models					Comparison of qualifications levels				
	Models groups	diff	lwr	upr	p adj	Levels groups	diff	lwr	upr	p adj
Testing data	RF—MLR	0.024	− 0.007	0.054	0.178	Level 2—Level 3	0.030	− 0.029	0.088	0.427
	SVM—MLR	0.021	− 0.010	0.052	0.282					
	XGBoost—MLR*	0.036	0.006	0.066	0.014	Level 1—Level 3	0.022	− 0.045	0.089	0.688
	SVM—RF	− 0.003	− 0.033	0.027	0.994					
	XGBoost—RF	0.0121	− 0.018	0.043	0.734	Level 1—Level 2	− 0.007	− 0.074	0.060	0.961
	XGBoost—SVM	0.015	− 0.015	0.046	0.578					
Estimating data	RF– MLR	0.027	− 0.010	0.064	0.222	Level 2—Level 3	0.028	− 0.061	0.118	0.674
	SVM—MLR	0.024	− 0.013	0.061	0.320					
	XGBoost—MLR*	0.044	0.006	0.081	0.017	Level 1—Level 3	0.026	− 0.078	0.129	0.783
	SVM-RF	− 0.003	− 0.040	0.034	0.996					
	XGBoost—RF	0.016	− 0.021	0.053	0.669	Level 1—Level 2	− 0.003	− 0.106	0.100	0.996
	XGBoost—SVM	0.019	− 0.018	0.057	0.535					

*$p < 0.05$; diff = the mean difference between two groups; lwr = the lower end point of the interval; upr = the upper end point of the interval; p adj. = the p value for the comparisons


Singapore
SMINS
ESCS
HOMEPOS
TMINS
ADINST
ST013Q01TA
PERFEED
TEACHSUP
WEALTH
ST004D01T
BELONG
EMOSUPS
ST011Q04TA
ST097Q03TA
FISCED
MISCED
HISCED
ST123Q02NA
Features
0 10 20 30
Importance (%)
Japan
SMINS
ESCS
TMINS
WEALTH
TEACHSUP
ST013Q01TA
PERFEED
ST097Q03TA
HOMEPOS
ADINST
BELONG
MISCED
ST004D01T
EMOSUPS
FISCED
ST011Q04TA
HISCED
ST123Q02NA
Features
0 5 10
Importance (%)
Estonia
ST013Q01TA
TMINS
ESCS
PERFEED
WEALTH
SMINS
TEACHSUP
ADINST
BELONG
ST097Q03TA
HOMEPOS
ST004D01T
EMOSUPS
FISCED
ST123Q02NA
HISCED
MISCED
ST011Q04TA
Features
0 5 10 15
Importance (%)
Chinese Taipei
ST013Q01TA
SMINS
TMINS
ESCS
WEALTH
PERFEED
HOMEPOS
ADINST
TEACHSUP
EMOSUPS
HISCED
BELONG
ST123Q02NA
ST097Q03TA
ST004D01T
FISCED
ST011Q04TA
MISCED
Features
0 5 10 15
Importance (%)
Finland
ST013Q01TA
SMINS
ESCS
PERFEED
ADINST
TMINS
WEALTH
TEACHSUP
BELONG
HOMEPOS
ST097Q03TA
MISCED
EMOSUPS
FISCED
HISCED
ST004D01T
ST123Q02NA
ST011Q04TA
Features
0 5 10 15 20
Importance (%)
Luxembourg
ST013Q01TA
ESCS
TMINS
SMINS
PERFEED
WEALTH
BELONG
ADINST
TEACHSUP
HOMEPOS
ST004D01T
ST097Q03TA
EMOSUPS
MISCED
FISCED
HISCED
ST011Q04TA
ST123Q02NA
Features
0 5 10 15 20
Importance (%)
Italy
ST013Q01TA
TMINS
PERFEED
SMINS
ESCS
ADINST
WEALTH
ST004D01T
TEACHSUP
HOMEPOS
BELONG
ST097Q03TA
HISCED
EMOSUPS
MISCED
FISCED
ST011Q04TA
ST123Q02NA
Features
0 5 10 15
Importance (%)
Hungary
ST013Q01TA
ESCS
TMINS
PERFEED
WEALTH
ADINST
HOMEPOS
BELONG
TEACHSUP
SMINS
ST097Q03TA
EMOSUPS
ST004D01T
MISCED
ST123Q02NA
HISCED
FISCED
ST011Q04TA
Features
0 5 10 15
Importance (%)
Lithuania
ESCS
ST013Q01TA
TEACHSUP
TMINS
PERFEED
ADINST
WEALTH
BELONG
ST097Q03TA
HOMEPOS
SMINS
EMOSUPS
ST123Q02NA
ST011Q04TA
FISCED
ST004D01T
MISCED
HISCED
Features
0 5 10 15 20
Importance (%)
Level - 3
Level - 2

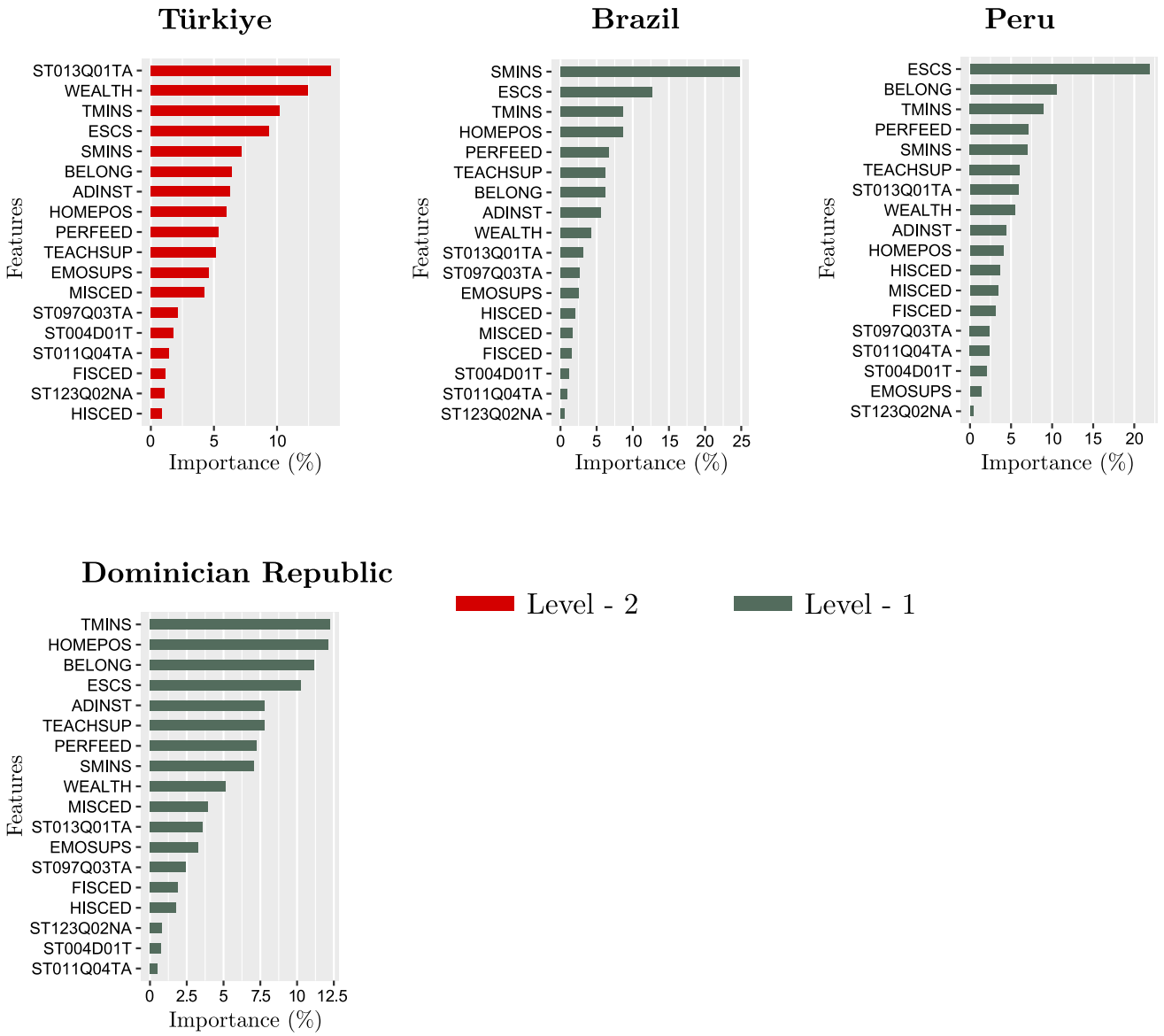


Fig. 8 continued

found to be the second most important variable (12.5%) only in Türkiye. Finally, according to the overall results summarized in Fig. 8, the difference between the variable with the highest significance level and the variable with the lowest significance was determined most in Singapore (28.9%) and least in Dominican Republic (11.9%).

3.7 Importance of PISA results: impacts on education systems

Presently, there is a huge amount of data collected about individuals in different environments and for different purposes and it is very important to determine the level of influence of these data in the decision-making process [27, 82–84]. Thanks to the international recognition that the PISA project has achieved and the prestige it has gained in the media, PISA results have become the most important indicator in many participating countries and formed the basis for education policies [8, 17, 31, 85–88]. It is understood from the increase in the number of countries participating in the exam that the interest in the PISA has increased over the years due to the results it reveals about the national education systems [46, 47].

Comparative evaluation of the findings obtained in PISA, implemented by the OECD, on the basis of common indicators, is important and meaningful for the education policies to be produced in this regard [1]. Researchers suggest that policymakers should take into account factors

affecting PISA results, review variables in detail, compare country educational performance, develop and regulate education programs in this context, and increase investment in education [6, 87–95]. For this reason, it is very important to determine the variables with high importance that affect the PISA results with scientific approaches (machine learning algorithms in this study) and for the countries to make improvements in their education systems based on these variables in order to achieve success.

4 Conclusions

In this study, PISA-2015 data of 13 randomly selected countries were trained separately with certain machine learning algorithms (MLR, SVR, RF, and XGBoost), and the models created at the end of the training, the PISA 2018 science achievement scores were estimated both as the individual score of the students who took the exam and as the average score of the countries participating in the exam. According to the findings obtained in the study, it has been determined that the XGBoost machine learning algorithm performs more accurate predictions by showing higher performance than the results obtained with other models in all countries investigated. It was determined that the highest PISA-2018 science achievement scores of the students who participated in the exam, estimated by this algorithm, were in Luxembourg ($r = 0.600$, RMSE = 75.06, MAE = 59.97), while the lowest were in Finland ($r = 0.467$, RMSE = 79.38, MAE = 63.24). In addition, the degree of correlation between the predicted scores and the actual scores of the students who took the exam in all of the researched countries was found at the level of moderate correlations ($0.40 < r < 0.69$). With the XGBoost algorithm, the average science scores of the countries were estimated, and the direction of change in the form of increase and decrease in the average science score as well as the average scores for all surveyed countries were estimated with high accuracy. This is an important piece of evidence showing that the future prediction performance of the machine learning model is accurate and high. In addition to these, the importance levels of the variables that affect the success of the model were also examined, and it was determined that the importance levels of the SMNS, TMINS, ESCS variables were high, and the FISCED, MISCED, HISCED, and ST123Q02NA variables were low in all of the countries studied. Additionally, in terms of qualification levels, no great differences were detected between the importance levels of the variables. As a result, according to all the findings obtained, it was revealed with this study that machine learning methods, and especially the XGBoost algorithm displayed a successful performance in predicting the science scores of the students who took the exam in the researched countries. Moreover, the future science achievement averages of the countries were estimated at the high-performance level with the same algorithms. In addition, the results revealed important findings that can be used in the development of educational strategies in the countries surveyed, especially considering the importance of the factors investigated.

Increasing the number of different and significant input variables that greatly affect science success, not using variables with low importance in success level in the machine learning process, and developing machine learning processes with hybrid methods can help further develop this study and make predictions with higher accuracy. However, it should be noted that the machine learning process and the prediction periods will be longer in that case. We recommend these issues be taken into consideration for future studies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-023-08901-6.

Acknowledgements The authors would like to express sincere gratitude towards OECD for its generous publication of the data.

Funding There is no funding for this study.

Data availability Data will be made available upon request.

Declarations

Consent statement Not applicable.

Conflict of interest The authors declare that they have no conflicts of interest. The authors alone are responsible for the content and writing of this paper.

References

1. Aydın A, Erdağ C, Taş N (2011) A comparative evaluation of pisa 2003–2006 results in reading literacy skills: an example of top-five OECD countries and Turkey. Educ Sci Theory Pract 11:665–673
2. PISA (2021) Programme for ınternational student assessment. https://www.oecd.org/pisa/. Accessed 25 Nov 2021
3. Aksu G (2018) PISA başarısını tahmin etmede kullanılan veri madenciliği yöntemlerinin incelenmesi (in Turkish). Dissertation, Hacettepe University, Ankara, Türkiye.
4. Hu J, Yu R (2021) The effects of ICT-based social media on adolescents' digital reading performance: a longitudinal study of PISA 2009, PISA 2012, PISA 2015 and PISA 2018. Comput Educ 175:104342. https://doi.org/10.1016/J.COMPEDU.2021.104342
5. Rolfe V (2021) Tailoring a measurement model of socioeconomic status: applying the alignment optimization method to 15 years of PISA. Int J Educ Res 106:101723. https://doi.org/10.1016/J.IJER.2020.101723

6. Gomes M, Hirata G, e Oliveira JBA (2020) Student composition in the PISA assessments: evidence from Brazil. Int J Educ Dev 79:102299. https://doi.org/10.1016/J.IJEDUDEV.2020.102299
7. Delprato M, Antequera G (2021) Public and private school efficiency and equity in Latin America: new evidence based on PISA for development. Int J Educ Dev 84:102404
8. Kim HJ, Yi P, Hong JI (2021) Are schools digitally inclusive for all? Profiles of school digital inclusion using PISA 2018. Comput Educ 170:104226. https://doi.org/10.1016/J.COMPEDU.2021.104226
9. Spruyt B, Van Droogenbroeck F, Van Den Borre L et al (2021) Teachers' perceived societal appreciation: PISA outcomes predict whether teachers feel valued in society. Int J Educ Res 109:101833. https://doi.org/10.1016/J.IJER.2021.101833
10. Odell B, Gierl M, Cutumisu M (2021) Testing measurement invariance of PISA 2015 mathematics, science, and ICT scales using the alignment method. Stud Educ Eval 68:100965. https://doi.org/10.1016/J.STUEDUC.2020.100965
11. 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 157:103964. https://doi.org/10.1016/J.COMPEDU.2020.103964
12. Song X, Mitnitski A, Cox J, Rockwood K (2004) Comparison of machine learning techniques with classical statistical models in predicting health outcomes. Stud Health Technol Inform 107:736–740. https://doi.org/10.3233/978-1-60750-949-3-736
13. Bühlmann P (2012) Bagging, boosting and ensemble methods. In: Gentle J, Härdle W, Mori Y (eds) Handbook of computational statistics. Springer, Berlin, pp 985–1022
14. Skurichina M, Duin RPW (2002) Bagging, boosting and the random subspace method for linear classifier. Pattern Anal Appl 5:121–135. https://doi.org/10.1007/s100440200011
15. Dietterich TG (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 402(40):139–157. https://doi.org/10.1023/A:1007607513941
16. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 143(14):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
17. Masci C, Johnes G, Agasisti T (2018) Student and school performance across countries: a machine learning approach. Eur J Oper Res 269:1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031
18. Chen J, Zhang Y, Wei Y, Hu J (2021) Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach. Res Sci Educ 51:129–158. https://doi.org/10.1007/s11165-019-9835-y
19. Pejic A, Molcer PS, Gulaci K (2021) Math proficiency prediction in computer-based international large-scale assessments using a multi-class machine learning model. İn: SISY 2021–IEEE 19th ınternational symposium on ıntelligent informatics, proceedings, pp 49–54. https://doi.org/10.1109/SISY52375.2021.9582522
20. Bozak A, Aybek EC (2020) Comparison of artificial neural networks and logistic regression analysis in PISA science literacy success prediction. Int J Contemp Educ Res 7:99–111. https://doi.org/10.33200/ijcer.693081
21. Lezhnina O, Kismihók G (2021) Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA. Int J Res Method Educ 45(2):180–199. https://doi.org/10.1080/1743727X.2021.1963226
22. Zeybekoglu Ş, Koğar H (2022) Investigation of variables explaining science literacy in PISA 2015 Turkey sample. J Meas Eval Educ Psychol 13:145–163. https://doi.org/10.21031/epod.798106
23. Saarela M, Yener B, Zaki MJ, Kärkkäinen T (2016) Predicting math performance from raw large-scale educational assessments data: a machine learning approach. In: JMLR workshop and conference proceedings, vol 48, pp 1–8. http://medianetlab.ee.ucla.edu/papers/ICMLWS3.pdf
24. Toprak E (2017) Yapay sinir ağı, karar ağaçları ve ayırma analizi yöntemleri ile PISA 2012 matematik başarılarının sınıflandırılma performanslarının karşılaştırılması (in Turkish). Dissertation, Hacettepe University, Ankara, Türkiye.
25. Aksu G, Doğan N (2018) Veri Madenciliğinde Kullanılan Öğrenme Yöntemlerinin Farklı Koşullar Altında Karşılaştırılması (in Turkish). Ankara Univ J Fac Educ Sci (JFES) 51:71–100. https://doi.org/10.30964/auebfd.464262
26. Hu X, Gong Y, Lai C, Leung FKS (2018) The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: a multilevel analysis. Comput Educ 125:1–13. https://doi.org/10.1016/j.compedu.2018.05.021
27. Bezek-Güre Ö, Kayri M, Erdogan F (2020) Analysis of factors effecting PISA 2015 mathematics literacy via educational data mining. Educ Sci 45:393–415. https://doi.org/10.15390/EB.2020.8477
28. Aksu N (2019) Farklı ülkelerden PISA sinavina katılan öğrencilerin matematik okur yazarlığını etkileyen faktörlerin tahmin edilmesi (in Turkish). Dissertation, Aydın Adnan Menderes University, Türkiye
29. Puah S (2020) Predicting students' academic performance: a comparison between traditional MLR and machine learning methods with PISA 2015. Dissertation, Ludwig-Maximilians-University, Munich,Germany. https://doi.org/10.31234/osf.io/2yshm
30. Rebai S, Ben Yahia F, Essid H (2020) A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socioecon Plann Sci 70:100724. https://doi.org/10.1016/j.seps.2019.06.009
31. Kjærnsli M, Lie S (2004) PISA and scientific literacy: Similarities and differences between the nordic countries. Scand J Educ Res 48:271–286. https://doi.org/10.1080/00313830410001695736
32. Erbaş KE (2005) Factors affecting scientific literacy of students in Turkey in Programme for International Student Assessment (PISA). Dissertation, Middle East Technical University, Ankara, Türkiye
33. Anıl D (2009) Factors effecting science achievement of science students in programme for international students' achievement (PISA) in Turkey. Educ Sci 34:87–100
34. Ho ESC (2010) Assessing the quality and equality of Hong Kong basic education results from PISA 2000+ to PISA 2006. Front Educ China 5:238–257. https://doi.org/10.1007/s11516-010-0016-z
35. Anagün Ş (2011) The impact of teaching-learning process variables to the students' scientific literacy levels based on PISA 2006 results. Educ Sci 36:84–102
36. Acar T, Öğretmen T (2012) Analysis of 2006 PISA science performance via multilevel statistical methods (in Turkish). Educ Sci 37:178–189
37. Sun L, Bradley KD, Akers K (2012) A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. Int J Sci Educ 34:2107–2125. https://doi.org/10.1080/09500693.2012.708063
38. Sälzer C, Heine JH (2016) Students' skipping behavior on truancy items and (school) subjects and its relation to test performance in PISA 2012. Int J Educ Dev 46:103–113. https://doi.org/10.1016/j.ijedudev.2015.10.009
39. Kahraman Ü, Çelik K (2017) Analysis of PISA 2012 results in terms of some variables. J Hum Sci 14:4797–4808. https://doi.org/10.14687/jhs.v14i4.5136

40. 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 40:273–295. https://doi.org/10.1080/01443410.2019.1652248
41. Karakoç-Alatli B (2020) Investigation of factors associated with science literacy performance of students by hierarchical linear modeling: PISA 2015 comparison of Turkey and Singapore. Educ Sci 45:17–49. https://doi.org/10.15390/EB.2020.8188
42. Rohatgi A, Scherer R (2020) Identifying profiles of students' school climate perceptions using PISA 2015 data. Large Scale Assess Educ. https://doi.org/10.1186/s40536-020-00083-0
43. Antonelli-Ponti M, Monticelli PF, Versuti FM et al (2021) Academic achievement and the effects of the student's learning context: a study on PISA data. Psico-USF 26:13–25. https://doi.org/10.1590/1413-82712021260102
44. Atasoy R, Çoban Ö, Yatağan M (2022) Effect of ıct use, parental support and student hindering on science achievement: evidence from Pisa 2018. J Learn Teach Digit Age 7:127–140. https://doi.org/10.53850/joltida.945869
45. Karakus M, Courtney M, Aydin H (2022) Understanding the academic achievement of the first- and second-generation immigrant students: a multi-level analysis of PISA 2018 data. https://doi.org/10.1007/s11092-022-09395-x
46. OECD (2015) OECD-PISA 2015 database. https://www.oecd.org/pisa/data/2015database/. Accessed 25 Nov 2021
47. OECD (2018) OECD-PISA 2018 database. https://www.oecd.org/pisa/data/2018database/. Accessed 25 Nov 2021
48. Vabalas A, Gowen E, Poliakoff E, Casson AJ (2019) Machine learning algorithm validation with a limited sample size. PLoS ONE 14:1–20. https://doi.org/10.1371/journal.pone.0224365
49. Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314. https://doi.org/10.1080/10618600.1996.10474713
50. R Core Team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.r-project.org/. Accessed 10 Oct 2022
51. Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab—an S4 package for kernel methods in R. J Stat Softw 11:1–20. https://doi.org/10.18637/jss.v011.i09
52. Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2/3:18–22. https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf
53. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: 22nd ACM SIGKDD ınternational conference on knowledge discovery and data mining, pp 785–794. https://doi.org/10.1145/2939672.2939785
54. Kuhn M (2008) Building predictive models in R using the caret package. Stat Softw 28(5):1–26. https://doi.org/10.18637/jss.v028.i05
55. Carslaw DC, Ropkins K (2012) Openair—an r package for air quality data analysis. Environ Model Softw 27–28:52–61. https://doi.org/10.1016/j.envsoft.2011.09.008
56. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
57. Pan S, Zheng Z, Guo Z, Luo H (2022) An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. J Pet Sci Eng 208:109520. https://doi.org/10.1016/j.petrol.2021.109520
58. Yeşilkanat CM, Kobya Y, Taşkın H, Çevik U (2017) Spatial interpolation and radiological mapping of ambient gamma dose rate by using artificial neural networks and fuzzy logic methods. J Environ Radioact 175–176:78–93. https://doi.org/10.1016/j.jenvrad.2017.04.015
59. Fu T, Tang X, Cai Z et al (2020) Correlation research of phase angle variation and coating performance by means of Pearson's correlation coefficient. Prog Org Coat 139:105459. https://doi.org/10.1016/j.porgcoat.2019.105459
60. Taylor R (1990) Interpretation of the correlation coefficient: a basic review. J Diagn Med Sonogr 6:35–39. https://doi.org/10.1177/875647939000600106
61. Kuhn M (2019) The caret Package. https://topepo.github.io/caret/index.html. Accessed 24 Mar 2023
62. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1111/j.1747-0285.2009.00840.x
63. Drucker H, Surges CJC, Kaufman L et al (1997) Support vector regression machines. Adv Neural Inf Process Syst 1:155–161
64. Ibrahem Ahmed Osman A, Najah Ahmed A, Chow MF et al (2021) Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Eng J 12:1545–1556. https://doi.org/10.1016/j.asej.2020.11.011
65. Wu J, Liu H, Wei G et al (2019) Flash flood forecasting using support vector regression model in a small mountainous catchment. Water (Switzerland) 11(7):1327. https://doi.org/10.3390/w11071327
66. Wu CL, Chau KW, Li YS (2008) River stage prediction based on a distributed support vector regression. J Hydrol 358:96–111. https://doi.org/10.1016/j.jhydrol.2008.05.028
67. Yu PS, Chen ST, Chang IF (2006) Support vector regression for real-time flood stage forecasting. J Hydrol 328:704–716. https://doi.org/10.1016/j.jhydrol.2006.01.021
68. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
69. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/BF00058655
70. Ho TK (1998) The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 20(8):832–844. https://doi.org/10.1109/34.709601
71. Yesilkanat CM (2020) Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. Chaos Solitons Fractals 140:110210. https://doi.org/10.1016/j.chaos.2020.110210
72. Wang H, Yilihamu Q, Yuan M et al (2020) Prediction models of soil heavy metal(loid)s concentration for agricultural land in Dongli: a comparison of regression and random forest. Ecol Indic. https://doi.org/10.1016/j.ecolind.2020.106801
73. Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199. https://doi.org/10.1007/s10021-005-0054-1
74. Gislason PO, Benediktsson JA, Sveinsson JR (2006) Random forests for land cover classification. Pattern Recognit Lett 27:294–300. https://doi.org/10.1016/j.patrec.2005.08.011
75. Zhang H, Yin S, Chen Y et al (2020) Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. J Clean Prod 273:122858. https://doi.org/10.1016/j.jclepro.2020.122858
76. Zeng N, Ren X, He H et al (2019) Estimating grassland aboveground biomass on the Tibetan Plateau using a random forest algorithm. Ecol Indic 102:479–487. https://doi.org/10.1016/j.ecolind.2019.02.023
77. Friedman J (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
78. Zhu X, Chu J, Wang K et al (2021) Prediction of rockhead using a hybrid N-XGBoost machine learning framework. J Rock Mech Geotech Eng 13:1231–1245. https://doi.org/10.1016/j.jrmge.2021.06.012
79. Ma M, Zhao G, He B et al (2021) XGBoost-based method for flash flood risk assessment. J Hydrol 598:126382. https://doi.org/10.1016/j.jhydrol.2021.126382
80. Hu L, Wang C, Ye Z, Wang S (2021) Estimating gaseous pollutants from bus emissions: a hybrid model based on GRU and

XGBoost. Sci Total Environ 783:146870. https://doi.org/10.1016/j.scitotenv.2021.146870
81. Wang J, He L, Lu X et al (2022) A full-coverage estimation of PM2.5 concentrations using a hybrid XGBoost-WD model and WRF-simulated meteorological fields in the Yangtze River Delta Urban Agglomeration, China. Environ Res. https://doi.org/10.1016/j.envres.2021.111799
82. Razzak MI, Imran M, Xu G (2020) Big data analytics for preventive medicine. Springer, London
83. Adi E, Anwar A, Baig Z, Zeadally S (2020) Machine learning and data analytics for the IoT. Neural Comput Appl 32:16205–16233. https://doi.org/10.1007/s00521-020-04874-y
84. Chen M, Mao S, Liu Y (2014) Big data: A survey. Mob Networks Appl 19:171–209. https://doi.org/10.1007/S11036-013-0489-0
85. Köhler O, Reiss K, Riecke-Baulecke T (2016) 15 Jahre PISA: Ergebnisse und Perspektiven (in German). Schulmanagemet-Handbuch, vol 157. Cornelsen Verlag, München
86. Çelen FK, Çelik A, Seferoğlu S (2011) Türk eğitim sistemi ve PISA sonuçları (in Turkish). In: XIII. Academic ınformatics conference, vol 765–773 https://yunus.hacettepe.edu.tr/∼sadi/yayin/AB11_Celen-Celik_Seferoglu_PISA-Sonuclari.pdf
87. Darling-Hammond L (2014) What can PISA tell us about US education policy? New Engl J Public Policy 26(1):1–14. https://scholarworks.umb.edu/nejpp/vol26/iss1/4/
88. Martens K, Niemann D (2013) When do numbers count? The differential impact of the PISA rating and ranking on education policy in Germany and the US. German Politics 22:314–332. https://doi.org/10.1080/09644008.2013.794455
89. Aydın A, Selvitopu A, Kaya M (2019) Eğitime Yapılan Yatırımlar ve PISA 2015 Sonuçları: Karşılaştırmalı Bir İnceleme (in Turkish). Elem Educ Online 17:1283–1301. https://doi.org/10.17051/ILKONLINE.2018.466346
90. Gürlen E, Demirkaya AS, Doğan N (2019) Uzmanların Pısa ve Tımms Sınavlarının Eğitim Politika ve Programlarına Etkisine İlişkin Görüşleri (in Turkish). Mehmet Akif Ersoy Univ J Educ Fac. https://doi.org/10.21764/MAEUEFD.599615
91. Dobbins M, Martens K (2011) Towards an education approach à la finlandaise? French education policy after PISA. J Edu Policy 27:23–43. https://doi.org/10.1080/02680939.2011.622413
92. Araujo L, Saltelli A, Schnepf S (2017) Do pisa data justify pisa-based education policy? Int J Comp Educ Dev 19:20–34. https://doi.org/10.1108/IJCED-12-2016-0023/FULL/PDF
93. Steiner-Khamsi G, Waldow F (2018) PISA for scandalisation, PISA for projection: the use of international large-scale assessments in education policy making—an introduction. Glob Soc Edu 16:557–565. https://doi.org/10.1080/14767724.2018.1531234
94. Lingard B, Sellar S (2013) Globalisation and sociology of education policy: the case of PISA. Contemp Debates Sociol Educ. https://doi.org/10.1057/9781137269881_2
95. Figazzolo L (2009) Impact of PISA 2006 on the education policy debate. Education International. https://www.ei-ie.org/file/474