UNIVERSITY OF TURKU

Modeling Academic Performance Using the PISA 2022 Database

Master's thesis
May 2025
Turku

Author:
ANDREEA STEFANIA NECULA

DEPARTMENT OF MATHEMATICS AND STATISTICS

Master's thesis

Subject: Data Science and Machine Lerning in Education
Author(s): Andreea Stefania Necula
Title: Modeling Academic Performance Using the PISA 2022 Database
Supervisor(s): Prof. Ion Petre
Number of pages: 78 pages
Date: 21.05.2025

The study aims to explore the predictive modeling of students' achievements using PISA 2022 dataset and directing attention to estimate scores in mathematics, reading and science. Moreover, the research wants to discover key features that influence directly student achievement, assessing the efficacy of some machine learning models trained on this dataset.

Key words: PISA 2022, predictive modeling, machine learning, student achievement

Table of Contents

Summary of tables

Summary of figures

Summary of acronyms

OECD – Organisation for Economic Co-operation and Development
PISA – Programme for International Student Assessment
ML – Machine Learning
LightGBM – Light Gradient Boosting Machine
XGBoost – Extreme Gradient Boosting
RAM – Random Access Memory
MAE – Mean Absolute Error
ESCS – Economical, Social and Cultural Status Index
HOMEPOS – Home Possessions Index
FISCED – Father's Highest Level of Education
MISCED – Mother's Highest Level of Education
HISCED – Highest Parental Education Level
KNN – K-Nearest Neighbors
MoEC – Ministry of Education and Culture
MSE – Mean Squared Error
GOSS – Gradient-based One-Side Sampling
EFB – Exclusive Feature Bundling
CatBoost – Categorical Boosting
HGBR – Histogram-Based Gradient Boosting Regressor
MAP – Maximum A Posteriori
IRT – Item Response Theory
USD – Unite States Dollar
HDI – Human Development Index
WLE – Weighted Likelihood Estimate
SPSS – Statistical Package for the Social Science

Chapter 1 – Problem statement and motivation

1.1Research Motivation and Goals

The study aims to explore the predictive modeling of students' achievements using PISA 2022 dataset and directing attention to estimate scores in mathematics, reading and science. Moreover, the research wants to discover key features that influence directly student achievement, assessing the efficacy of some machine learning models trained on this dataset.

Education is the main factor in shaping the future for a child and impacting national education policies. A well-educated person with knowledge, critical thinking and skills may contribute to a prosperous society [0]. Moreover, education provides opportunities for their professional growth and self-fulfillment. PISA test, the main tool for evaluating educational systems from various countries by analyzing students aged 15 results in mathematics, reading and science by utilizing information about their home and school environments. Led by the OECD, this project, PISA assessment, yields valuable insights into different educational systems [1].

This programme is focused on measuring students' competences in mathematics, reading and science applied in real-world contexts. The dataset PISA 2022 includes substantial information about student achievements, socio-economic background, factors within the educational context that facilitate an analysis of the factors impacting academic performance [1].

The main goal is to find the better model or a good combination of predictive models that assess students; scores in mathematics, reading and science using the features available in the dataset. Moreover, estimating these scores with a high accuracy can support the detection of the most important features that contribute to scholar success.

This objective is supported by machine learning models that aim to extract these characteristics to improve the academic environment and provide personalized guidance to students facing academic difficulties. In this project, we will use both ML models based on decision trees and linear regression models including Simple Decision Tree Regressor, Random Forest Regressor, LightGBM, XGBoost Regressor, Gradient Boosting Regressor, Histogram Gradient Boosting Regressor, Ridge and Lasso Regression, ElasticNet Regression, to evaluate which models offer better prediction accuracy on our dataset.

The academic purpose of this project is the assessment of the results obtained for each trained model. This will allow drawing some conclusions about model sensitivity, parameter tuning to prevent overfitting and model stability. Additionally, another aspect is reducing training time through appropriate optimizations. The dataset is quite large, containing approximately 60,000 samples, meaning that each model may require a significant amount of time to process all the data, depending on the RAM resources. The PISA test includes 81 countries, but the goal is to finish a specific analysis for a selected number of countries and to observe how students from different nations are influenced differently in their academic trajectories [1].

Moreover, the results of this project may be a valuable resource for governments or teachers and researchers who seek to improve educational systems and develop solutions for the needs of each student. Furthermore, this study seeks to predict students' score in Mathematics, Reading and Science for some selected countries: Singapore, Finland, Romania, Bulgaria, Estonia, Germany, France,

Hungary and Serbia. Additionally, the research focuses on identifying the most influential features that leads these predictions, giving insights into the key contributions to student achievement. Furthermore, in recent years, educational systems are turning more and more data based approaches to identify the key factors behind student succes. This research investigates the predictive modeling of student achievement using the PISA 2022 dataset with a particlar focus on mathematics, reading and science. What's more, this study follows which factors are the most significant and to see which machine learning models give the most accurate results across 9 countries, with no overfitting. Also, this project is grounded in the idea that education plays a key role in the growth of the both individuals and society as a whole. A strong educational background enchaces not only critical thinking and cognitive skills, but also a ability to participate actively in a modern economy [66]. PISA coordinated by the OECD plays an important role in offering a global benchmark bt testing how well 15 years olds solve everyday problems [66].

Machine learning models, espeacially ensemble models offer and effective alternative to traditional statistical techniques. Also, tree based models such as XGBoost and Random Forest perform much better than linear models [63]. Similarly, Huang et al. [64] used SHAP based interpretatuons of XGBoost to reveal the most important features related to mathematical literarcy such as student confidence and digital access. Moreover, Öz E. [65] proposed an ensemble method that works across different countries. This algorithm indicating that combining models makes prections more reliable to various education systems. This reflects the approach used in this study which involves training and testing 10 regression models on 9 different countries using tree based and linear models. This study also follows the good practices recommended by the OECD for working with the PISA database, helping to ensure accurate and repeatable results [67]. What's more, the main goal is to suport teachers, and researchers by helping them to identify students risk early on, ajust educational strategies and improve teaching methods based on solid data. By predicting student scores and analysing trends in each country the finding may support future efforts to customize new educational strategies.

1.2. Applicability

The study presented above provides valuable insights into educational policy, school management, student support systems and the social environment of each student. Moreover, good predictions for students' performance can help identify potential issues in their academic trajectories. Additionally, early identification of struggling students, allows interventions in time for academic changes such as: resource allocation, optimization of teaching methods and the development of personalized learning programs. At the same time, an important application of this research is personalized education that involves pinpointing the key factors that primarily influence students. As a result, we can introduce adaptive learning systems, designed to provide students with tailored theoretical and practical resources that address to their specific academic needs. These applications can be developed at the national level, based on individual analysis. On top of that, the study proposes insights into minimizing educational disparities. A good example is the identification of socio-economic barriers that impede student's success, consequently decreasing educational inequality. Beyond that, schools with lower PISA scores could have additional funding for modernization, as well as for teachers training courses to improve their teaching techniques and familiarize themselves with new personalized teaching methods. Additionally, curriculum designers can use the insights gained from

this study to adapt and create new courses that support students and teachers, ensuring continuous efficiency in the learning process.
In conclusion, we can affirm that this study contributes to the development of new educational programs focused on helping students who face learning difficulties, as well as shaping social factors that can enhance the quality of student learning through predictive modeling and analysis of obtained results.

1.3 Layout of the thesis

This study follows a systematic structure to analyze how environment, number of digital with screen at home or the number of books at home could improve or affect children's academic performance using advanced mathematical modeling approaches and machine learning models. Below is a detailed breakdown of each chapter.
Chapter 1 - Problem statement and motivation: This chapter show the motivation behind the study, the main topics of the project. Also, it discusses the importance of using regression models for analyzing academic performance and justifies the choice of PISA 2022 dataset.Furthermore, this chapter show the aplicability of the results in educational field to improve student outcomes.
Chapter 2 – Overview of prior studies: This section provides an overview of relevant studies about regression models used in educational analysis, with a particular emphasis on previous studies that use the PISA Dataset for performance prediction, highlighting different approaches that have been used within the field of education.
Chapter 3 – Machine learning architectures: Provides fundamental information needed to understand the techniques used in educational data analysis. It covers tree-based models, linear regression models and a discussion about feature ranking and adjustments for model performance.
Chapter 4 – Data preprocessing and analysis: This chapter explores the initial steps in preparing the dataset. It concentrates on data cleaning, handling missing values, scaling and normalizing.
Chapter 5 – Experimental results: This chapter covers the technical details such as explanation of how models were set up for experimentation. Also, the results from tree-based and linear models are highlighted in this chapter. Furthermore, this section includes an analysis of how model performance varies for different countries and some comparisons of results obtained through different models.
Chapter 6 – Analysis of model performance across countries: This capter presents the performance of ten regression models trained on nine country specific subsets across the three domains assessed in PISA: mathematics, reading and science. Additionally, the to 10 most important features identified for each country and domain are displayed in Figures 6.2 through 6.10, offering the key features.
Chapter 7 – Conclusions and Future work: This chapter summarizes the findings of the study and show the limitations. Additionally, it explores different techniques that could improve the predictions in future research.Also, the final part summarizes the effectiveness of the various models and draws findings derived from models results.

Chapter 2 – Overview of prior studies

2.1 Regression models used in educational analysis

Commonly, in educational research, regression models are frequently used to analyze and predict student performance and the factors which influence it. These models help us to construct an idea about the relationship between students, school factors and the academic outcomes in a statistical manner. Regression models are used to identify key factors in student performance such as socio-economic factors, school resources, home digital resources and quality of teaching. Using this dataset, PISA 2022, the regression models can perform a wider analysis with a higher accuracy because of large and diverse dataset, including 81 countries. Also, in PISA 2022 edition, nearly 700.000 students from 81 countries have participated [8].

Recent studies have explored the power of machine learning in educational field, trying to extract the most important features for student academic performance, based on PISA 2015 and PISA 2018 datasets. The research used data from 13 countries randomly selected. Also, the study compared multiple tree-based regression models, such as: Support Vector Machines, Extreme Gradient Boosting Regressor and Random Forest Regressor. Additionally, the data was split 75% training set and 25% validation set. XGBoost has recorded the best performance on the data selected. This model recorded the highest R2 score and MAE for Luxemburg with a coefficient of determination equal with 0.60, whereas Finland recorded the minimum score for performance, 0.467. Furthermore, this research highlights the most important features such as time spent studying science (SMINS), socio-economic and cultural background (ESCS), home resources (HOMEPOS) and parental education levels (FISCED, MISCED, HISCED) [4].

The application of machine learning techniques in education assessment was made in several studies. Another research demonstrated that regression-based models could successfully predict student math performance handling at the same time the missing values. The study exposes the advantages of feature extraction in reducing computational complexity but maintaining the accuracy and the stability of the model [6].

To sum it up, the mix between regression models and machine learning techniques reveals a good performance in predicting academic performance and features extraction. Moreover, the traditional methods such as linear regression might remain useful for understanding the relationships between variables.

2.2 Previous studies on PISA 2022 data analysis

The PISA (Programme for International Student Assessment) initiative is a comprehensive dataset for educational analysis, measuring performance of the models at the same time. The variety of the dataset and approach multiple studies in academic field. From feature ranking and the most important variables for each country, to find the correlation between several factors involved in research. A comparative study between 13 countries was made by Masci and his teams, who wanted to assess the key factors in student's performance using machine learning methodologies. These results highlighted the importance of school, the quality of teachers, the contribution of digital devices in the student's academic life, the number of books at home and more. So, different variables play an important role in students' education and they may differ from one country to another [4]. Another study applied machine learning in the educational context trying to estimate the life satisfaction of UK and Japanese

secondary students using PISA 2018. The report applied Random Forest Regressor (RF) and k-Nearest Neighbors (KNN), making use of the UK and Japan subsets from the PISA 2018 dataset. This study measured students' life satisfactions from UK and Japan leveraging 26 key features from PISA 2018 dataset such as digital resources, teacher support, bullying and sense of meaning in life. The Random Forest Regressor models performed better than K-NN, having a better accuracy. Also, both models performed better on UK students' subset than Japanese students' subset. The most important feature found in students' life satisfactions were sense of purpose, teacher assistance, exposure to bullying, and digital devices in both home and school environments [5]. The article [7] emphases how the PISA results were applied in Finnish education policies from 2000 to 2006. The researchers were analyses the programs by the Ministry of Education and Culture to see what the PISA assessment influenced regarding education. Furthermore, the article shows that Finnish authorities used the results of PISA test to justify the educational mechanism and to promote this type internationally. But after the Finland's decline from 2012, where the scores decreased, the MoEC started finding the issues from the educational field, applying minor changes. So, the PISA test was used as a main pillar justifying the actual educational policies and promoting their educational plan. Despite the initial success, students' scores have been decreasing at PISA test highlights question marks over the academic rules [7].

2.3 Challenges in predicting student performance

Analyzing the student performance based on PISA 2022 may be a complex task. Numerous factors contribute to the output. Having a large data set could be problematic sometimes. Filling the missing values represent an important step in preprocessing the data, the time of execution, feature selection, data complexity also could be a barrier in a project like this.

In the study [5], they have identified a major problem being the missing values. This significant issue hinders to have a high accuracy. Also, the models performed better on the UK student subset than on Japanese student subset. This discovery suggests us that the cultural differences may affect performance of the models being hard to generalize on unseen data from another country with a different educational and cultural background [5]. Another study by Saarela (2016) explored the prediction of mathematics performance based on large-scale educational assessment datasets. The main topic discussed in this article was high-dimensional data and finding a solution for this. With thousands of features for every student, running the models may be complicated and a wasting of resources. For this reason, the feature selection was essential, reducing computational costs while keeping a high accuracy. The author solved the problem, applying unsupervised learning technique to extract the key predictive feature before using models for prediction [6]. While the previous author focused on optimizing the model's performance-wise and saving resources, Seppänen (2019) highlighted the difficulties of using the statistical results of PISA test in educational field. Instead of using just the ranking, they could develop deeper analyses in educational systems [7].

Chapter 3: Machine learning architectures

This field represents a subchapter within artificial intelligence that enables computers to recognize patterns from large datasets. It permits computers to extract patterns from input data and then it can make informed predictions and decisions, or some classification without requiring specific commands. Instead of traditional programs, machine learning models have a good performance because they are exposed to more data [8]. These important topics make ML a powerful tool in classification, regression and pattern recognition. Also, it can be used in a variety of domains such as examined in detail through this study. The methodologies of machine learning may be grouped into three fundamental types. The first one, the supervised learning, involves training models with data points that have labels. Labeled dates are the inputs which have a known output. So, the models that use this approach learn the mapping between input and output variables, then applied on unseen data to make new predictions. Commonly it incorporates linear models and various artificial neural networks. These kinds of ML systems are used in diverse applications in various fields enumerated above [9]. Unsupervised learning handles unlabeled data. The aim is to find patterns in data and group them by similarity. The clustering is not the only procedure that unsupervised learning could do. It can handle high-dimensional data to a low-dimensional grid, where similar data points are placed closed to each other. Several approaches, such as dimensionality reduction methods like Principal Component Analysis and clustering approaches such as K-means and Hierarchical Clustering, are utilized [9]. An essential aspect of reinforcement learning is that an agent learns to optimize its decisions and performs at his tasks by interacting with the environment, then receiving rewards if it did the best choice. There are also penalties for agent does not make a good decision. After a while the agent improves its skills optimizing his strategy and maximizing the total rewards. This type of learning is used in robotics or game playing [9].
In this research the focus will on supervised ML algorithms including tree-based models and linear models, comparing their performance. The focus of this project will be on education, assessing the performance of the models on PISA 2022 dataset. Furthermore, despite the advantages of machine learning multiple challenges data quality, missing values, overfitting and computational resources will appear.

3.1. Tree-based models

3.1.1 Simple Decision Tree Regressor

Decision trees represent a machine learning models applied for classification and regression problems. Also, this model is a non-parametric supervised learning algorithm. A Simple Decision Tree Regressor contains a root node where there are the whole datasets, then internal nodes that include some features for splitting the data, branches which represent the decision which is made based on features and the leaves which show the class in the case of classification and numerical prediction in case of regression problems. In addition, this type of models works recursively trying to split the dataset into subsets based on some condition at every step, developing a hierarchical-tree composition. The main goal when you build a decision tree is to find the smallest tree, meaning the minimum numbers of nodes that fits the training dataset without any errors. The length of a decision tree may grow until a certain rule is met such as: the maximum depth or the minimum impurity is reached, several examples from a node fall below a threshold. On top of that, choosing the optimal

split involves some metrics to measure the error [10]. Another important aspect is the overfitting phenomenon. It appears in the model when the tree becomes too complex, fitting the data very well, but cannot generalize on unseen data. Therefore, to enhance the model's efficiency, pruning method is utilized. The last mentioned is a process of reducing the size of the tree, stopping the overfitting. Pruning may be identified as pre-pruning or post-pruning. The first one limits tree growth before will be too complex, conditioned by maximum depth, minimum number of examples or minimum impurity threshold, the method being called early stopping. The post-pruning is about removing the redundant nodes or which not improve the performance after the tree is fully built.

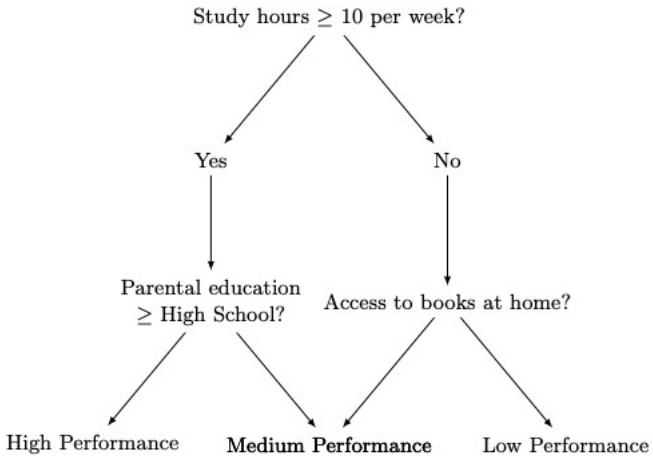


Figure 3.1 Simple Decision Tree Regressor

3.1.2. Random Forest Regressor

This model uses decision trees to make predictions by aggregating multiple models. This methodology is applied in both classification and regression tasks, building multiple decision trees and then combine the predictions to return an output. Also, in regression tasks, the output is generated using the mean value of prediction and for classification problems the output represents the most frequent class among the predictions of all models in the ensemble [12]. Moreover, this type of algorithm may reduce the overfitting, making the prediction on unseen data more accurate and stable. These benefits appear because an ensemble model can generalize better than o Simple Decision Tree Regressor [13].

Bagging, or bootstrap aggregating, is a machine learning approach (technique) that combines several models from the same category to return a more stable and accurate prediction. Each decision tree is built using a subset from the original dataset [14]. As previously mentioned, regression outputs are determined by averaging predictions, while classification outputs are based on the most frequent class [12]. Additionally, model performance can be evaluated using Gini Impurity for partitioning tasks and Mean Squared Error (MSE) for regression problems [14].

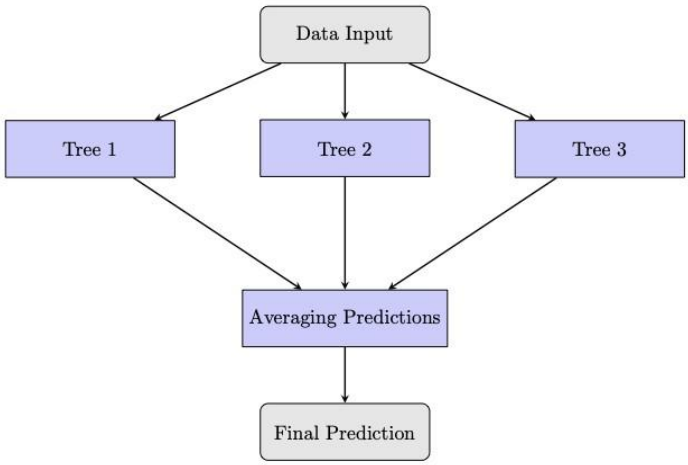


Figure 3.2: Ensemble Tree Regression

3.1.3. Gradient Boosting Regressor

This model represents a powerful ensemble learning technique that strengthens regression model performance by mixing several models which do not perform extraordinary, commonly decision trees. Also, this technique develops models progressively, with each new learning model from the mistakes of the previous ones, which increases accuracy. This method was introduced by Friedman, who illustrated the success of the algorithm in handling predictive analytics [23]. Furthermore, Gradient Boosting Regressor manages the errors by enhancing new models which approximates the opposite slope of the objective function from the present ensemble. The framework, at iteration m is formally defined as follows in formula:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \quad (1)$$

where the parameter $F_{m-1}(x)$ expresses the previously developed model also, $h_m(x)$ highlights the weak learner added at the iteration m, while ν signifies the learning rate which controls how much $h_m(x)$ play a role in the ensemble [24].

Therefore, the algorithm continues the process until a requirement is satisfied such as number of trees is reached, or the loss function reaches the convergence. This model is focused on optimizing the differentiable loss functions, allowing it to become highly significant for diverse set of regression problems. Also, to prevent the overfitting, Gradient Boosting Regressor uses valuable strategy such as shrinkage and subsampling. The shrinkage decreases the influence of every tree adjusting its contribution by a scaling factor ν. The subsampling technique highlights the training on a random subset from the initial dataset for each tree. This method helps the model to avoid the overfitting, generalizing better on unseen data [25]. This model was applied over multiple fields such as education, health, engineering and more. There is a study where Gradient Boosting Regressor was implemented in cheminformatics to predict and analyze molecular properties. Emphasizing the robustness in chemical analysis, the researchers laid out a set of instructions for training, optimizing and evaluate the model for molecular property prediction [26]. According to another study the algorithm mentioned was tested in modeling diabetes onset using extensive datasets. It proved that the framework works very well on medical data with complex relationships between points and disease outcomes [27]. Furthermore, an important study in neurorobotics demonstrates the value of Gradient Boosting Regressor applied to enhance the accuracy of robot movement performance. Thus,

continuously improving its predictions, the model accomplished more accurate trajectory estimations than old methods for prediction [28].

Figure 3.3 Gradient-Boosted Regression Tree

3.1.4. Light Gradient Boosting Machine Regressor

LightGBM is a new version of gradient boosting architecture built in 2016 by Microsoft. It is applicable to both classification and regression tasks. In its previous version, Gradient Boosting Regressor has some techniques which grow the tree length, but the new version LightGBM, employs a grow strategy based on leaves. Moreover, this model is a histogram-based decision tree learning algorithm. In comparison with the traditional gradient boosting algorithms, LightGBM groups the feature values into discrete bins and then develops histograms which approximate the data distribution. The effect of binning process is dramatically the reduction of memory usage and computational complexity, giving a faster model without compromising the accuracy. Additionally, to optimize further the performance, it introduces two new concepts including the Gradient-Oriented One-Sided Sampling (GOSS) and Optimized Feature Bundling (EFB). The method known as GOSS performs efficiently on big datasets, where prioritizes data points with high gradient values. This method is useful because the data points with larger gradients have valuable information about the errors from the model. Another important aspect about GOSS is that it does not sample randomly the data points. Instead of sampling randomly, it keeps the meaningful points with high gradients and select randomly from those with low gradients. This approach ensures a high accuracy, reducing the number of samples and a strong focus on the key components of the dataset. Conversely, the EFB method handle the issue of computational overhead is caused by high-dimensional data. In numerous datasets there are features which contain a significant number of zeros and have just few nonzero values. In consequents, EFB mixes those features that aren't together, minimizing the number of the features. So, this arrangement is arranged by a greedy algorithm that approximate the best combination, helping the model to run faster on large datasets, being accurate and stable [15].
The study by Anghel (2018) compared some machine learning algorithms which built decision trees to make predictions. The research reveals that the above-mentioned model is faster than other models such as XGBoost and CatBost. The reason of this performance is that LightGBM grows its trees in a different way, picking the best part of the tree to extend, instead of extending all the branches in the same way. So, this kind of strategy allows it to work faster, especially for large datasets [16].
According to Florek and Zagdański (2023) who compared the LightGBM, XGBoost and CatBoost performance, the first one trains the data faster than the other tested models, keeping a good prediction. They highlighted that while all models had a good accuracy, LightGBM is superior through the handling the large and complex datasets. This important property makes it recommended in the cases where the computational performance is a major criterion [17]. A recent investigation by Salvador (2024) analyzed the poverty levels in households in Philippines. In this study the best performance having the LightGBM. The model was able to group the families into different economic categories from the initial dataset. The finding may confirm the multiple fields where the model could

be used. Not only in technical field, but also in social science, helping governments or personal in charge to develop new applications improving the quality of life [18].
This is further supported by the mathematical formulation of LightGBM. This algorithm derives from Gradient Boosting Regressor, where every new tree undergoes training to decrease its error from previous trees. The mathematical mission is to optimize the function L (y, F(x)), where the parameter y represents the target label and F(x) represents the predicted output. This model is updated at each iteration using the formula:

$$F_{\$}(x) = F_{\$"\#}(x) + \eta f_{\$}(x) \quad (2)$$

where the learning rate (η), which adjusts the contribution of each new tree plays a role in the final prediction [35]. Also, to find a new optimal tree, this model decreases the error by trying to adjust the negative gradient of residuals, formula:

$$f_{\$} = \arg\min_{\%} \Sigma_{\&(\#}^{'} L4y_{\&}, F_{\$"\#}(x_{\&}) + f(x_{\&})7 \quad (3)$$

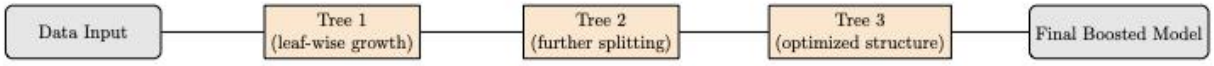


Figure 3.4 Light Gradient Boosting Machine Regressor

3.1.5. Extreme Gradient Boosting Regressor

Extreme Gradient Boosting Regressor came from Gradient Boosting Regressor framework's family. It is an optimized version that is developed to ensure a strong performance in terms of efficiency and accuracy. XGBoost is known for the capacity to process large datasets and keeping an exceptional performance. In comparison with traditional tree-based models, XGBoost uses a smart gradient boosting method in a parallelizable manner, increasing computational efficiency and improving the accuracy. Moreover, for regression analysis, this model develops an ensemble of trees, the main task being the error reduction from antecedent tree. Also, this algorithm adopts mean squared error function to measure the performance in regression problems [19]. A mathematical approach is highlighted in formula:

$$\mathcal{L}^{(t)} = \Sigma_{\&(\#}^{'} l :y_{\&}, \dot{y}_{,}^{(\$"\#)} + f_{\$}(x_{\&})< + \Omega(f_{\$}) \quad (4)$$

where the parameter t represents the current iteration, the variable l corresponds with the loss function, yi symbolizes the actual variable and yi-1 the prior one, ft(xi) corresponds to the new tree appended to suppress the error and the parameter Ω(ft) represents a penalty function which prevent the overfitting [19]. Furthermore, major advantage of this model is the capacity of handling the missing values from the datasets. Instead of filling the missing values manually, the model finds the best way to deal with them while assembling the tree. Therefore, this key capability increases its applicability on real-world datasets, where missing values is a common problem. Additionally, another skill for XGBoost represents the usage of first order and second derivatives (gradient and Hessian) to find the best modality to split the points. To sum it up, the models do not consider just the gradient (loss function), but also its curvature (Hessian), constructing the model more robust and computationally efficient during the training because of the second-order Taylor approximation [19]. The main idea is highlighted by formula:

$$L(y, F_{\$"\#} + f_{\$}) \approx L(y, F_{\$"\#}) + g_{\&} f_{\$}(x_{\&}) + \frac{\#}{-} h_{\&} f_{\$}^{-}(x_{\&}) \quad (5)$$

A recent study [20] highlighted the proficiency of the model trying to predict the static and dynamic Young's modulus of sedimentary rocks developed from physical and structural attributes. The algorithm obtained great values for performance's parameters, such as R2= 0.997 for static and R2=0.999 for dynamic modulus, on complex dataset [20]. Additionally, another study revealed the high performance of the model trained on a biomass gasification system. In this research was analyzed the impact of equivalence ratio and heating value. The result, R2=0.96 showed a good prediction, suggesting the XGBoost's consistency in engineering field and in energy field [21].
Another study investigates the performance of XGBoost in the financial field. It was used for predictions in cryptocurrency prices. Also, in this study there were introduced parameters which allow the identification of patterns in cryptocurrency movements. The analysis confirms that the model fits very well non-linear relationships between data, becoming a valuable instrument for financial field [22].

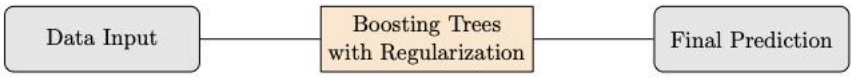


Figure 3.5 Extreme Gradient Boosting Regressor

3.1.6. Histogram – Based Gradient Boosting Regressor

The Histogram Gradient Boosting Regressor comes from the Gradient Boosting models' family, but an advanced algorithm which can deal with large datasets.
It uses depth-wise tree growth, exploring all branches at the same time before trying to go deeper. This procedure makes the algorithms more stable, but unfortunately slightly less efficient than LightGBM [29]. In comparison LightGBM approaches leaf-wise growth strategy, symbolizing that expand the leaves first then reduces the loss function. It is a good practice for a faster convergence, but it gives with a high probability of overfitting [15]. Also, the Histogram Gradient Boosting Regressor converts the continuous features into discrete bins, aiding in the formation of the histogram which approximate the data distribution. As a result, this approach reduces the memory usage and computational overhead, obtaining a faster training procedure and a high accuracy [15]. The training phase contains the development of histograms for each feature, where continuous feature values are grouped into several bins. Instead of trying to assess each possible value, the model finds the best split on these histograms. Therefore, this strategy minimizes the tome of finding the best split from O(n*log(n)) to O(nbins), where the parameter n represents how many samples are, while nbins is how many bins are, smaller than n. To sum it up, the training is more efficient, with a focus on large datasets. The mathematical approach is highlighted in formula:

$$F_{\$}(x) = F_{\$"\#}(x) + \eta \cdot h_{\$}(x) \quad (6)$$

where the parameter $F_{t-1}(x)$ represents the ensemble models from previous iteration, $h_t(x)$ suggests the new decision tree added at current iteration and the learning rate (η), the contribution of ht(x) at the final output. Thus, the split based on the binned data rather than the raw continuous data, highlights an efficient computation [29]. Moreover, the key strength of HGBR is its ability to naturally handle missing values. During training, the algorithm identifies the optimal path to the appropriate child node for samples with missing values for every split by evaluating the potential information gain. Thus, this method allows the model to incorporate missing data directly without requiring prior imputation [30].

HGBR has been applied successfully across multiple fields. For example, in finance, it has been used to predict stock prices for Alphabet Inc., achieving high accuracy and demonstrating its capability in capturing complex financial patterns [30].
In another study, for environmental science, HGBR has been employed to estimate groundwater mean residence times. By analyzing hydrogeological characteristics, the model provided precise predictions of groundwater ages, contributing to better water resource management [31].
Overall, the Histogram-Based Gradient Boosting Regressor presents an efficient and reliable solution for regression problems, particularly with large datasets and high-dimensional features. Its approach, based on feature binning and histogram-based split selection, improves computational efficiency without compromising accuracy, making it an asset in modern machine learning [15].

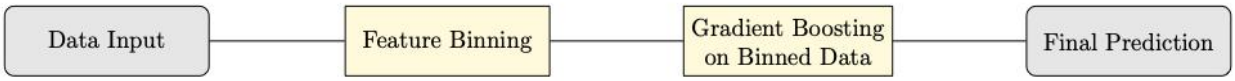


Figure 3.6 Histogram-Based Gradient Boosting Regressor

3.2. Linear Models

3.2.1. Linear Regression

Linear regression is the most common model adopted to understand the relationship between some variables. It shapes the correlations among dependent variable and multiple independent variables presuming a linear connection between them [32, 33]. Thus, a mathematical representation for this algorithm could be illustrates by formula:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (7)$$

where the parameter Y represents the depended variable while X is the independent one. Also, the variable β0 is the value of Y while X is zero and β1 signifies the coefficient of X and ϵ expresses the error [34, 35]. For multiple linear regression the formula has a new structure:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (8)$$

because multiple predictors are recognized. This algorithm was used in several fields including social science, economy, medicine and engineering [36].
Furthermore, estimation's parameters linear regression uses the Linear Regression strategy that requires the discovery of the metrics for β0 and β1 for the purpose of reducing the sum of squared residuals, optimizing model fit to the data. This concept is illustrated below in formula:

$$\beta_1 = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2} \quad (9)$$

where the parameter y_i represents the observed values, and y_i characterizes the predicted output from the regression equation. Also, the estimated parameters are portraited in formula (9) and formula:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad (10)$$

where $\bar{X}$ and $\bar{Y}$ represent the average values of X and Y [35]. Thus, this estimation strategy is a good practice to decrease the difference between the target value and the real output, defining a robust approach for this algorithm.
In addition, to ensure that a linear regression model offers good predictions, a few essential demands must be accomplished. Firstly, there should be a linear connection between the dependent variables and independent ones. If this condition is not met, transformations or alternative modeling approaches

should be explored. Secondly, the observations must be independent, meaning that the dependent variable's values should not be correlated across different instances. Also, the residuals should maintain homoscedasticity that meaning their variance should remain consistent across all layers of the independent variable X. Lastly, the residuals should respect a standard distribution. If these principles are not met, the model may produce biased estimates, making it essential to verify these conditions before using linear regression [34, 37]. On the top of that, this algorithm is commonly applied across various disciplines. For instance, in economics is used to evaluate relationships such as consumer spending relative to income or the link between inflation and unemployment [35,36]. Also, linear regression can be seen in medical research, it serves as a tool for predicting disease risk based on patient characteristics, such as assessing the effect of blood pressure on cardiovascular conditions [37]. Additionally, in engineering, linear regression helps analyze system performance and forecast sensor readings, providing valuable insights into how different factors affect operational efficiency [36]. In conclusion, linear regression is a fundamental technique in statistical modeling and predictive analysis. The structure of the model is simple, but offers good and valid results, and clear relationships between data [32].

3.2.2. Ridge Regularization

This model is employed to deal with multicollinearity in linear regression. Multicollinearity appears when the independent variables became highly correlated, making the design matrix X nearly singular and this fact induces to unstable estimates in Linear Regression. Thus, to tackle this problem, Ridge Regression adds a new regularization parameter which shrinks the regression coefficients, minimizing the variance and enhancing the stability of the model [38]. The Ridge Regression estimator is constructed by introducing penalties to the OLS loss function framework. This procedure is illustrated in formula:

$$L(\beta) = |Y - X\beta|^{-} + \lambda|\beta|^{-} \quad (11)$$

where the Y factor stands for the respond vector, the design matrix by X, the coefficient vector by β and the regularization factor by λ. Additionally, $\lambda\|\beta\|2$ improves a better generalization by decreasing the overfitting, because the $\|\beta\|2$ is the Euclidian distance, and the λ parameter limits the overall magnitude of coefficients [39]. Moreover, an important benefit of this model is that can offer a unique solution even where XT*X is singular, showed in formula:

$$\beta = (X^{/}X + \lambda I)^{"\#}X^{/}Y \quad (12)$$

where the XT parameter is the transpose of the design matrix [40]. The addition of λ*I guarantee that the matrix inversion is always possible, thus stabilizing the regression estimates. Also, having an appropriate regularization parameter is important, as it determines the contrast between bias and variance. A larger λ increases bias but decreases variance, helping to improve the predictive performance, especially in cases where multicollinearity is present. On top of that, techniques such as cross-validation are commonly used to find the best regularization parameter which minimizes the loss function [41]. Alternatively, from a Bayesian perspective, the framework may be analyzed as a maximum a posteriori estimator. This aspect is possible if the regression coefficients lead an independent normal prior with mean zero and common variance. This interpretation provides a probabilistic framework for understanding how Ridge Regression shrinks coefficients based on the prior assumption [42]. Ridge Regression represents a powerful tool for handling multicollinearity within linear regression frameworks, adding a regularization term, scaled by the total of the squared

parameter values. This is known as L2 regularization because it uses the L2 norm of the coefficient vector [38].

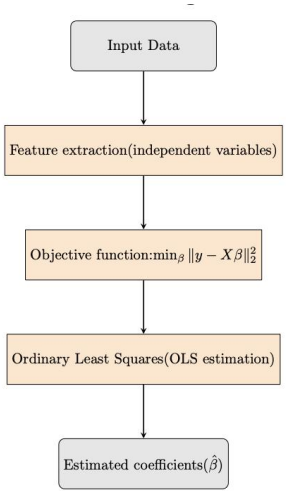


Figure 3.7 Linear Regression

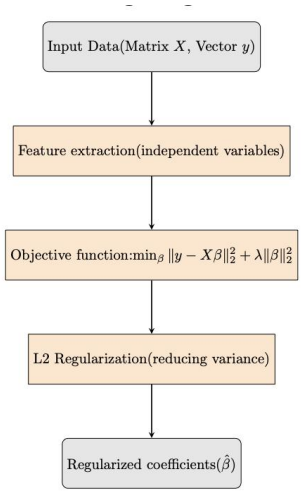


Figure 3.8 Ridge Regression

3.2.3. Lasso Regression

Lasso represents a predictive technique designed to improve model performance by introducing L1 regularization, which reduces the number of nonzero coefficients. The standard least-squares regression optimizes just the total squared deviations, but Lasso Regression adds a constraint term proportional to the total absolute magnitude of the coefficients. This procedure changes specific weights being scaled down to exactly zero, choosing important variables and improving the prediction of the model [43]. Moreover, the Lasso Regression approach is illustrated in formula:

$$\hat{\beta} = \arg\min_{\beta_0} \left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \quad (13)$$

where the yi parameter represents the target variable, xij is the features, β0 represents the intercept, βj denotes the regression coefficients and λ acts as the regularization factor modeling penalty constraints applied to the total magnitude of the coefficient values [44]. This approach L1 regularization term $\lambda\sum|\beta j|$ forces some coefficients to become exactly zero, as a result, the algorithm keeps just the most important features [45]. Choosing an optimal λ value is essential because it determines the balance between model complexity and prediction accuracy. A higher λ leads to more coefficients being set to zero, simplifying the model but potentially increasing bias. Also, a lower λ retains more variables, reducing bias but possibly increasing variance and overfitting. To select an optimal λ, cross-validation techniques are commonly used, ensuring that the model generalizes well to unseen data [46]. A major advantage for Lasso Regression is its ability to handle high-dimensional datasets where the number of predictors exceeds the sample size. This makes it particularly useful across genetics, finance, and text analysis, where selecting relevant variables from a large set is necessary [47]. Compared to Ridge Regression, which also applies regularization but retains all predictors, Lasso provides a sparse solution, reducing the risk of overfitting. [48]. From a Bayesian perspective, Lasso Regression corresponds to the MAP estimate under a Laplace prior on the regression coefficients. This assumption justifies the sparsity-inducing nature of Lasso, as the Laplace prior encourages smaller coefficients, shrinking some of them to exactly zero [49].

Lasso Regression is a powerful alternative to OLS regression when dealing with high-dimensional datasets or when variable selection is necessary. By introducing an L1 penalty, it shrinks coefficients and removes irrelevant variables, leading to better generalization and interpretability. Selecting the appropriate λ is critical to balancing model complexity and predictive accuracy, making cross-validation an essential part of the Lasso modeling process [43].

3.2.4. Elastic Net Regression

The Elastic Net Regression model represents a blend between Lasso and Ridge Regression models that improve the efficiency for high-dimensional datasets. The framework was introduced to resolve the limitations of Lasso Regression, which struggles with selecting one variable from a group of highly correlated predictors, and Ridge Regression, which does not perform variable selection. Thus, combining L1 and L2 regularization, Elastic Net Regression maintains a balance between selecting the important features and minimizing the coefficient size [50]. Ridge Regression reduces this problem by applying L2 regularization, which reduces all coefficient magnitudes without setting any to zero. Lasso Regression, in contrast, utilizes an L1 regularization, that forces some coefficients to zero, choosing the most relevant features. However, Lasso often selects only one variable from a group of highly correlated predictors, which may lead to suboptimal models. Elastic Net Regression addresses this by combining L1 and L2 penalties, enabling the selection of correlated variables together while it is still promoted sparsity [51]. Elastic Net Regression problem is represented as an optimization function that minimizes the residual sum of squares while applying L1 and L2 penalties, formula:

$$\hat{\beta} = \arg\min_{0} \{||y - X\beta||^{2} + \lambda_{1}||\beta||_{1} + \lambda_{2}||\beta||^{2}\} \quad (14)$$

where the parameters λ_1 and λ_2 determines how strongly the L1 and L2 penalties affect the model. So, a larger lambda1 makes more coefficients shrink to zero, creating a simple model with fewer variables. Also, a larger lambda2 keeps all predictors but makes their values smaller, helping to prevent the overfitting phenomenon without removing any variables completely [54]. Elastic Net regression proves effective in scenarios when there are more features than samples, such as genomic studies, where thousands of genetic markers may be potential predictors of a disease. Since these markers tend to be highly correlated, standard Lasso Regression would select only a subset of them, ignoring relevant variables. Similarly, in finance, Elastic Net Regression is often used. The asset prices show strong dependencies, requiring a model that can mix multiple related features [55]. One of the key challenges in Elastic Net Regression is selecting optimal values for λ_1 and λ_2. These parameters are typically determined using cross-validation, where different values are tested, and the combination that produces the lowest prediction error is chosen. This approach ensures that the model generalizes well on unseen data while keeping only the most important variables while removing unnecessary ones. The Elastic Net approach is also closely linked to Bayesian statistics, as it can be viewed as MAP estimator with Bayesian prior that combines the Laplace and Gaussian distributions, reflecting the dual nature of the L1 and L2 penalties [56]. To sum it up, ElasticNet regression provides a powerful alternative to traditional regularized regression methods by blending the advantages of Lasso and Ridge. Its ability to perform variable selection while retaining correlated predictors makes it a flexible and effective technique for handling high-dimensional datasets. By adjusting the balance between L1 and L2 penalties, Elastic Net allows practitioners to fine-tune models based on specific data characteristics, improving both interpretability and predictive performance [50].

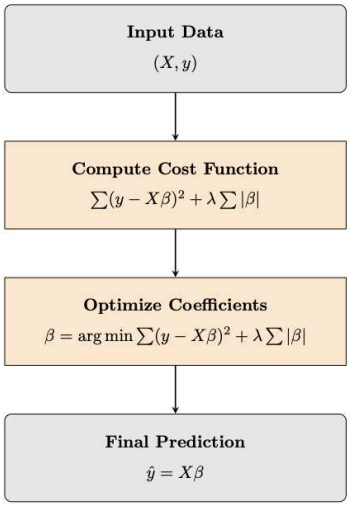


Figure 3.9 Lasso Regression

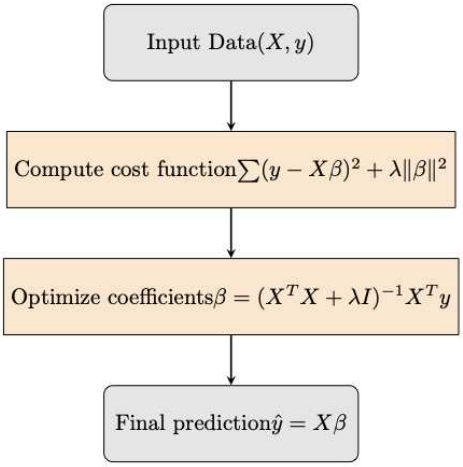


Figure 3.10 Elastic Net Regression

Chapter 4 - Data preprocessing and analysis

4.1 PISA2022 overview

The Programme for International Student Assessment represents a three-yearly international assessment undertaken by Organisation for Economic Co-operation and Development. It appraises the performance of students aged 15 in the areas of mathematics, literacy and science. This test is centered on their abilities to apply the knowledges on real-world scenarios rather than memorizing formulas and applying them in a mechanical way in exercises [55]. The main objective involves analyzing student performance in different countries with different economic situation, revealing the most effective educational frameworks and methodologies which may help others learning systems [56]. Launched in 2000, PISA has experienced new participants' countries every year. Therefore, in the 2022 assessment, 81 countries were part of this study, comprising around 690.000 students worldwide. Thus, this large-scale involvement supports a comprehensive analysis of global educational systems, allowing nations to make comparison between students' performance and adopts new practices from one another's experiences [57]. Moreover, the PISA examination was mainly based on mathematics, assessing the students' skills in logical thinking and problem-solving. As result, reading and science were minor domains, while creative thinking was introduced to measure the ability to create original ideas. Additionally, financial literacy was included as an optional module acknowledging the growing relevance of financial skills in our society today [58].

The assessment process involves two stages to extract the samples. Firstly, schools are randomly selected to embody the 15-years-old student population across the country. Then, a random group of students are selected to participate. This type of sampling ensures the diversity, having different educational contexts [59]. After the pandemic period, PISA 2022 included the data from all participants even though some countries had low response rate. This decision aimed to develop a complex analysis of the pandemic's impact over the global education systems. Thus, the results of PISA 2022 illustrate a paramount image over the landscape of education [60].

The assessment includes three major domains: mathematics, reading and science. In each cycle, one of them is prioritized and others two tested as minors. In PISA 2022, the main subject was mathematics which underscores the fact that more questions are associated with mathematics performance [57]. Each student has a targeted selection of questions. The assessment is computer based and lasts two hours. It contains multiple-choice questions, open-ended responses and interactive problem-solving tasks. The questions from the mathematic section evaluate the ability to interpret, formulate and solve mathematical problems, including real-life contexts such as financial calculus, geometric problems and data analysis. Additionally, the exercises have graphs, tables, even interactive tools. These test questions measure competencies such as estimating distances, measurements in daily life, interpreting graphs and using mathematical models to solve practical problems. Reading comprehension questions assesses how well the student understands and interprets texts such as blogs, newspapers, scientific articles or fictional stories. Also, the main requirements involve extracting key information from a paragraph or identifying the author's perspective. The science part tests students' ability scientific knowledge in real situations, including fields like biology, chemistry or physics. This part of the test verifies the understanding of the scientific methods, applying scientific reasoning to everyday problems. In 2022, apart from the three core subjects, PISA 2022 introduced new components that measure additional competencies. One of them is the creative

thinking which evaluates the ability to generate new ideas or to improve existing solutions. The questions covered topic such as writing a short story focused on given prompts or solving puzzles that require unconventional thinking. Another added topic is financial literacy which tests the understanding of financial concepts such as budgeting, financial risks or loans. The assessment featured a mix of multiple-choice questions, short-answer responses, extended responses or interactive digital tasks. Furthermore, the set of questions varies from one student to another. PISA uses a test booklet design, which means each student completes a different set of questions extracted from many items. It employs a method known as Item Response Theory (IRT) to ensure the comparability of results among students, even if they receive a different set of questions. Also, the questions are organized into multiple blocks, and they are randomly distributed across students' test booklets. Also, each cycle includes new developed questions and items from previous assessments, measuring the trends over time. Alongside the academic test, students, teachers and school administrators filled the background questionnaires which collected data on socio-economic status like family income or parents 'education level, school resources such as availability of technology or number of teachers, study habits, learning environments or students' well-being and mental health. This procedure helps the researchers to correlate the educational outputs with social and economic factors, revealing what influences student success. The test does not provide an individual score for each student but rather generates an aggregated national score to compare educational systems across different countries. The score is reported on a 1000-point scale, but the most scores are between 300 and 600 points. The average score for OECD countries is around 500 points for each domain. The OECD is an international institution that promotes politics aimed at improving the economic and social development of its member countries. The Organisation for Economic Co-operation and Development includes 38 countries. Although PISA is organized by OECD, participation is not limited for the member's countries. Thus, each test cycle includes OECD countries and non-member countries. In 2022 edition 81 countries participated, but just 38 OECD members. Also, they analyze separately the average score of its member. As a result, the OECD tends to score higher than non-OECD participants as evidenced in this study. Additionally, the Table 4.1. provides a list of countries analyzed in this study. The column "country" lists the countries included in the research, "code" represents the country code, "GDP per capita (USD)" indicates the average economic output per person. Also, "HDI" is Human Development Index which highlights the country's level of development, and the last column specifies which country is a member of the Organisation for Economic Co-operation and Development.

Country	Code	GDP per capita (USD)	HDI	OECD member	Number of samples
Singapore	SGP	82,503	0.939	No	6606
Finland	FIN	53,983	0.938	Yes	10239
Romania	ROU	14,861	0.821	No	7364
Bulgaria	BGR	12,622	0.816	No	6107
Estonia	EST	26,379	0.892	Yes	6392
Germany	DEU	58,150	0.942	Yes	6116
France	FRA	50,729	0.903	Yes	6770
Hungary	HUN	18,729	0.854	Yes	6198
Serbia	SRB	9,218	0.802	No	6413

Table 4.1 Study country data

4.2 Data preprocessing

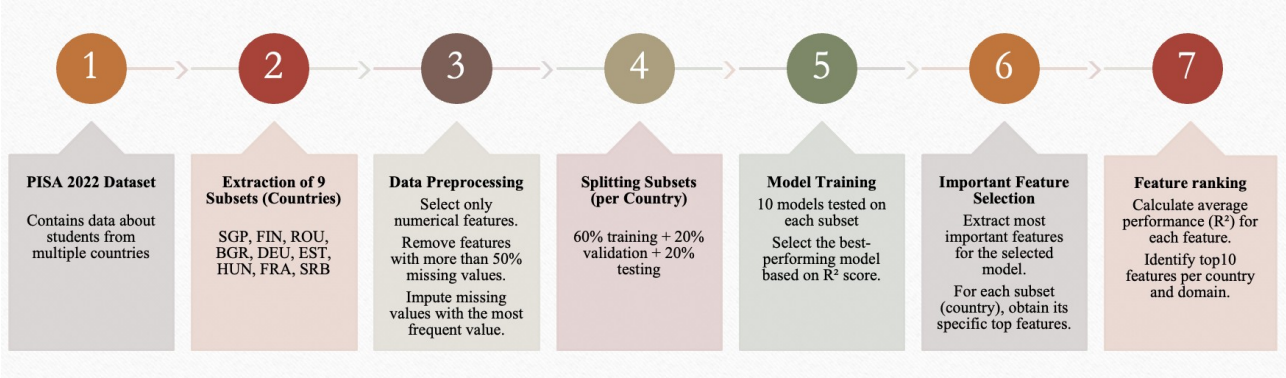


Figure 4.1 Project steps for PISA 2022

The dataset includes a wide range of student's performance information such as socio-economic background and school performance metrics. Thus, to ensure the consistency of the data, missing values were replaced with the most frequent imputation, which handle the Nan values with the most commonly value in that feature. Additionally, for preprocessing, was applied Min-Max Scaling to standardize and normalize numerical features, ensuring that all values are in the range [0,1], displayed in Appendix 2. This transformation improves the model performance, especially for the algorithms sensitive to feature magnitude. Moreover, for feature selection, only numerical features were retained, while categorical, non-informative and WLE features were removed since they didn't contribute directly to prediction. Additionally, the PValues for mathematics, reading and science were excluded from the input because they represent the target values. Including them in the input set, they would introduce a data leakage, leading to artificially high model accuracy. In the Pisa dataset, PValues represent the plausible values for student performance in measured competencies. The plausible values are generating using statistical estimation method called multiple imputation. The student performance cannot be measured with absolute precision due the factors such as test structure, sampling variability and student engagement. Instead of assigning a single fixed score to each student, PISA offers multiple possible values, 10 per subject. These values represent the range of scores a student would have achieved if they had answered a larger set of questions. In addition, the main purpose of the plausible values is to reduce measurement error. Thus, in data analysis, the plausible values should not be treated as independent values but rather as representations of the same fundamental performance estimation. Also, for complex analysis such as machine learning predictions, it is important to handle carefully the PValues to avoid bias in the results. These values in the input set would lead to data leakage because contain several data that we aim to predict. Moreover, to evaluate the model's performance, R2, coefficient of determination, and mean squared error were used as a key metrics. The coefficient of determination highlights how well the independent variables explain the changes in the target values. Measurements closer to the 1 indicate a good prediction and a lower value, suggests that the model does not learn enough from the data.
The mean squared error represents the average squared differences between target and predicted values, a lower MSE indicating a precise model. Together, the metrics assess the effectiveness of the model in predicting the students' performance. Firstly, each subset contains 1278 features. However, some features have many missing values, which leads to an unrealistic extraction of the most important features. Therefore, before training the selected models, I performed a data analysis to determine the percentage of missing vales for each feature. This step each repeated for each subset.

Moreover, following the analysis, it was observed that over 60% of the features contain more than 50% missing values. As a result, I decided to impute missing values using the most frequent value for those features with less than 50% missing data and to eliminate the rest. Leaving variables with no real information, just imputed from a single value, could cause them to appear important by mistake, during the feature importance extraction process, as they introduce noise. Furthermore, the method used to determine feature importance, permutation importance, assumes that a feature has a meaningful distribution in relation to the target variable. The assumption does not hold when data is artificially filled in through various imputation methods. Also, the distribution of missing values for each subset may be observed in Table 4.2, which details the number of features with less the 50% missing values, between 50% and 80% and more than 80%.

Country	<50%	50% - 80%	>80%
SGP	543	135	406
FIN	475	267	342
ROU	475	264	345
BGR	557	275	252
EST	558	182	344
DEU	454	417	213
FRA	410	266	408
HUN	675	245	164
SRB	373	265	446

Table 4.2 Missing data overview

An important step in building and testing machine learning models is to split the available data. This step ensures that the performance is evaluated is a fair way. Also, the data is usually divided into three parts. The training set used to teach the model, the validation set used to tune and compare models and the test set used to evaluate the final model. In this project the data was split as follows: 60% for training, 20% for validation and 20 % for testing. First, 20% of the data was kept for testing. The remaining 80% was split again, using 75% for training and 25% for validation. Therefore, this method ensures the model has enough data to learn from, while we were able to improve the models based on validation set, keeping unused the testing set. Also, a random seed was used. By setting this seen, we make sure that every time the code is run, we got the same data split and the same initial conditions. This fact ensures, a consistency in results, making easier the comparison between experiments. In general validation of machine learning models is used to assess the expected error of a learner on unseen datasets. So, each learner is optimized individually on the training dataset and their expected error are compared. What's more, validation sets are used to train the competing learners and evaluate their performance before selecting the final model. Also, the test set is only used to estimate the expected error of the final model, after all training and validation steps are complete. Speaking about the error on the training set, it is always smaller than the true error, which is why test sets are crucial for evaluation. In this study, comparing several models using unseen data allows us to identify which model generalize better [11].

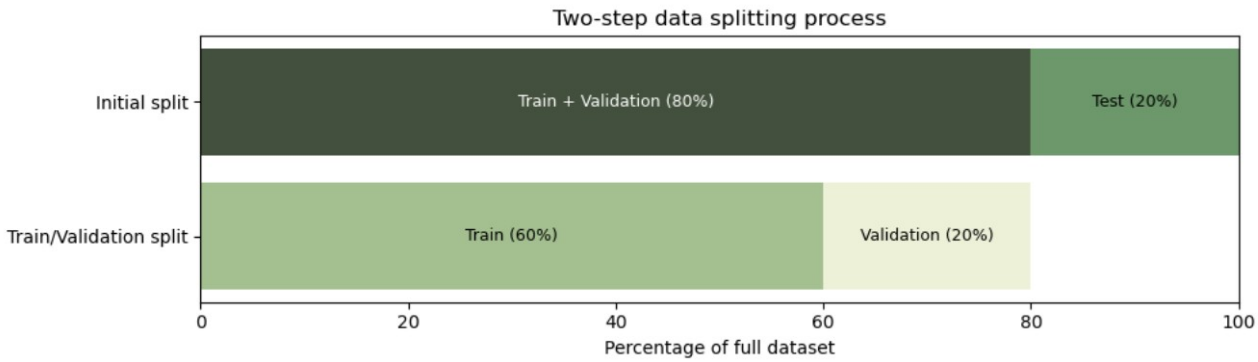


Figure 4.2 Data splitting scheme

4.3 Feature extraction

In this research, feature importance ranking represents an essential step to identify the most important factors which influence the students' performance in math, reading and science. This method used is called permutation importance. This technique evaluates the contribution of each feature by measuring how much the model's performance decrease or increase when the feature's value is randomly shuffled. Thus, this approach furnishes a clear understanding of which features are important and which not. Since the main target is to compare features across different countries, the ranking process is extended to provide a country specific analysis. So, for each country's model, the top 10 features with the highest permutation importance are selected for each domain. What is more, the values assigned by permutation importance are normalized so that they are comparable across models. This step is critical, because raw importance values may differ significantly based on dataset characteristics or model behavior. Once the importance values are normalized, the average importance for each feature is computed. This procedure help determine which features are consistently influence the students' performance prediction. Additionally, following this method, we can analyze which certain factors such as: education, school resources or study habits, have an impact on students' performance. Moreover, this feature ranking may help to improve education quality through the student performance in the PISA dataset.

For a better understand about the subsets and the student performance across different countries, I explored the distributions of mathematics, reading and science scores for each country. This study highlights how the spread of the data can influence the result of regression's models. Thus, for each I plotted a histogram and overlaid the mode and mean to observe the shape and the distribution of the data. The majority countries from the study present approximately normal distributions, with notable asymmetry in its distribution.

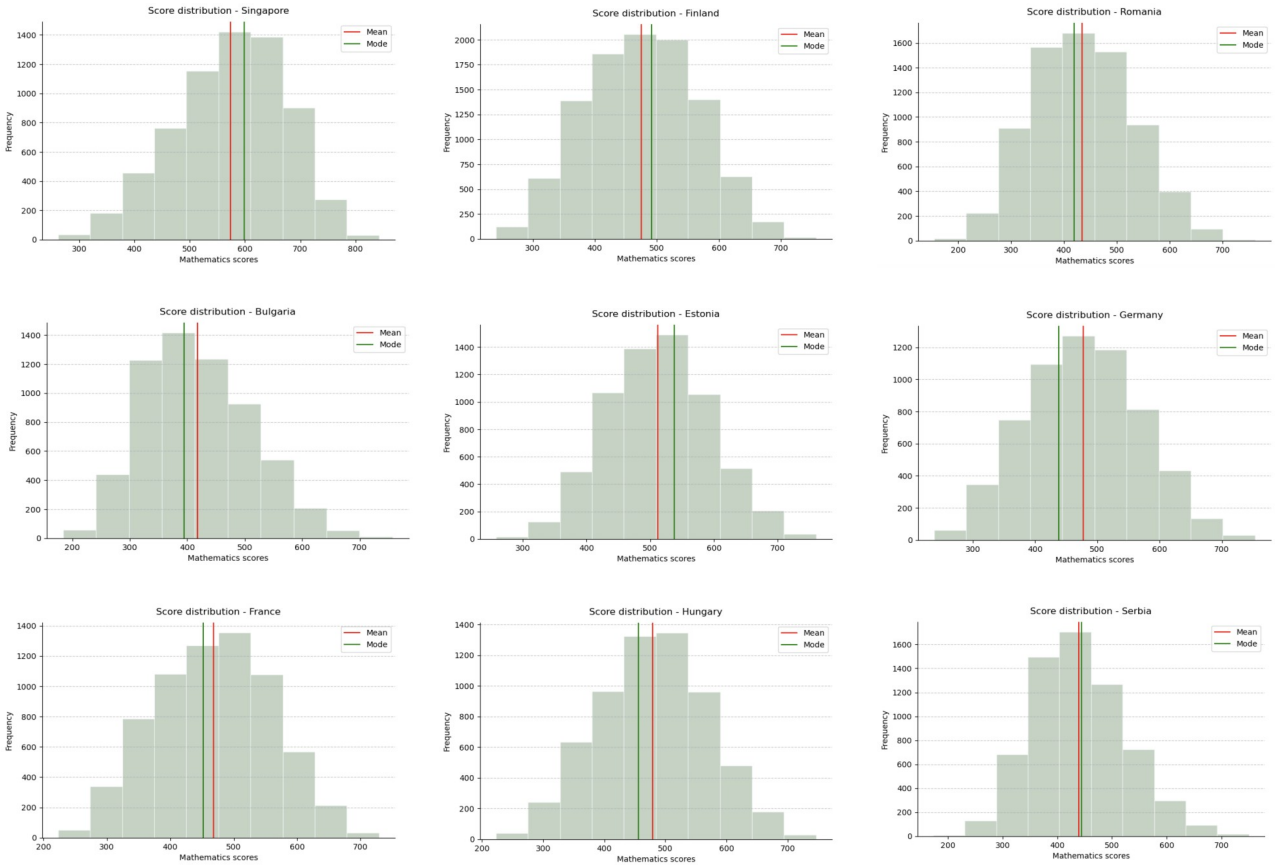


Figure 4.3 Distribution of students' mathematics scores by country

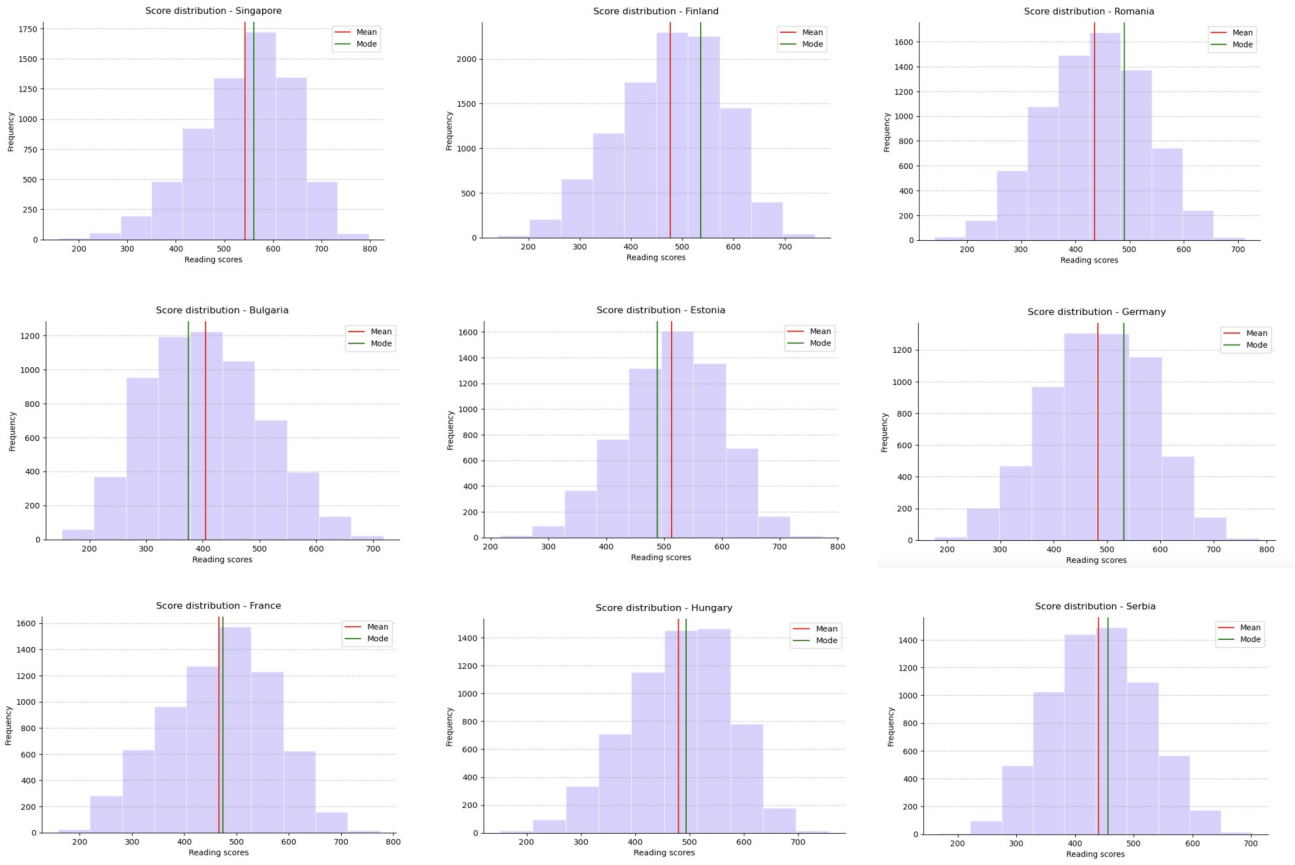


Figure 4.4 Distribution of students' reading scores by country

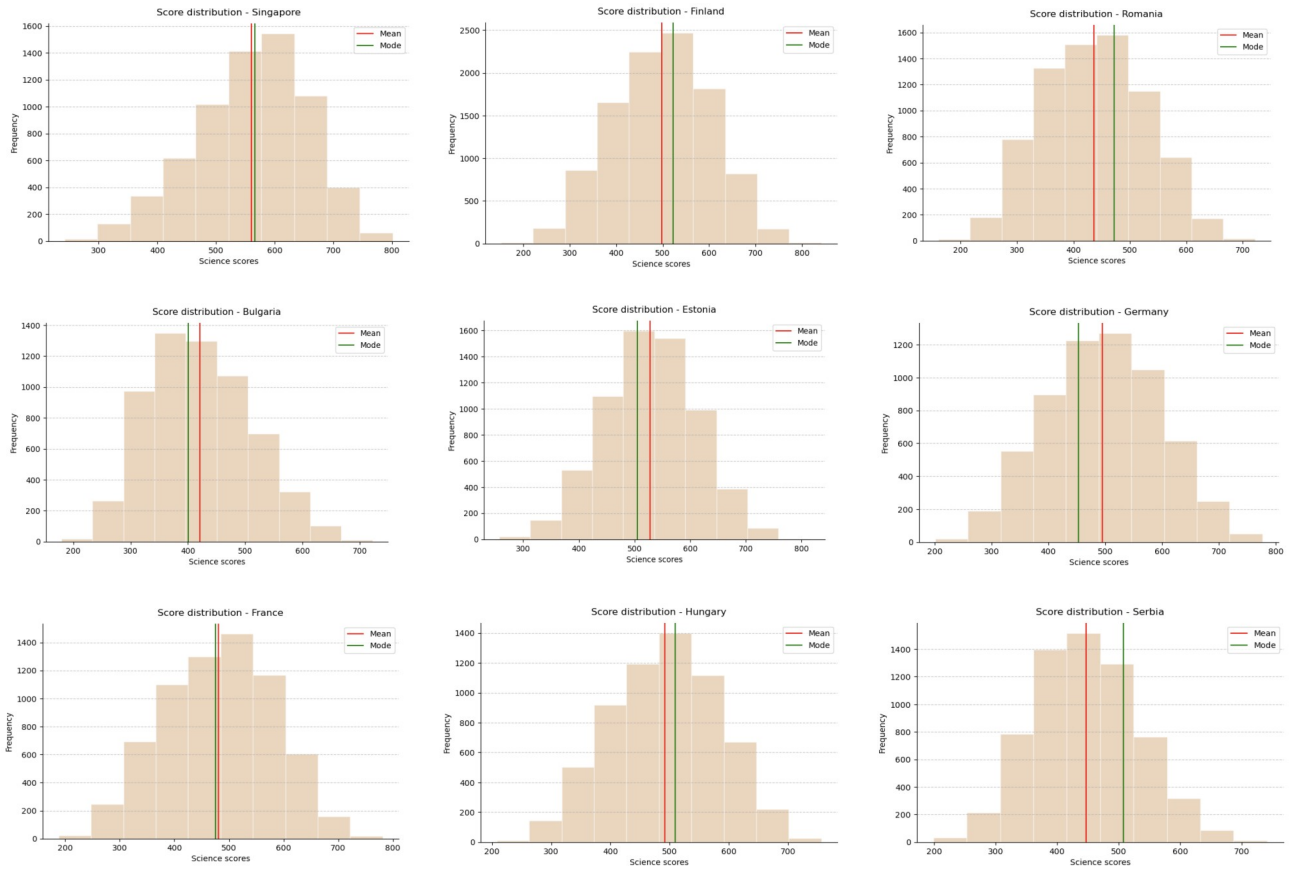


Figure 4.5 Distribution of students' science scores by country

Moreover, for each subset, was computed the mean, mode and standard deviation of students' score. Each subset includes three domains: math, reading and science and the statistical measures were calculated individually for each domain. The results are summarized in Table 4.3, where each row represents a domain within a specific country. These descriptive statistics provide an overview of the central tendency and variability of students' performance across domains and countries.

Country	Domain	Mean	Mode	Std
SGP	math	573.48	599.00	99.34
	read	541.98	561.00	99.14
	science	560.49	566.00	93.92
FIN	math	474.83	491.00	88.44
	read	477.03	536.00	101.83
	science	498.02	522.00	104.53
ROU	math	435.05	419.00	93.46
	read	435.86	490.00	93.51
	science	436.28	472.00	90.97
BGR	math	417.84	394.00	91.81
	read	404.92	374.00	100.04
	science	421.22	401.00	88.16
EST	math	512.21	538.00	80.30
	read	513.35	488.00	84.27
	science	528.02	505.00	82.38
FRA	math	467.75	452.00	90.89

	read	466.51	474.00	103.34
	science	480.90	475.00	100.01
DEU	math	477.24	438.00	90.19
	read	482.54	531.00	98.94
	science	495.03	453.00	100.27
HUN	math	479.23	456.00	88.52
	read	479.48	493.00	93.42
	science	491.94	509.00	90.42
SRB	math	439.32	444.00	83.91
	read	439.84	456.00	83.06
	science	446.80	508.00	83.81

Table 4.3 Statistical interpretation

4.4 PValues

In this research, I used the plausible values provided by the PISA 2022 dataset to estimate students performance in mathematics, reading and science. Furthermore, each student does not receive a single test score for reading, math, or science. Instead, they are given a set of ten scores called plausible values (PV1 to PV10). These are not exact results but rather possible estimates of a student's ability. They are based on how the student answered the test questions [66]. Therefore, to make the data easier to work with, I calculated the average of the 10 values for each domain. This gave me one stable score for each student in math, reading and science. This method gave me a better and more complete view of each student's abilities. By using these average scores, the models I trained had more consistent inputs, which helped improve prediction quality. Overall, working with these PVs was an important part of the project. This gave me a reliable way to compare students across different countries and supported more accurate machine learning analysis later in the study.

Because students in PISA only answer a part of the total test (due to time limits), it is not possible to know their real ability exactly. To solve this, researchers use statistical models to predict a range of reasonable scores each student could have. These predictions are the plausible values. They help us better estimate average performance at the national or group level, without making incorrect assumptions [68].

Also, for the PISA 2022 dataset analysis, we cannot use only one plausible value for each student. In my project, I calculated the averag of the 10 plausible values for each student in each domain: mathematics, reading and science and I I used these averages as the target in the regression models.

This method ensures that our findings are accurate and not biased. Also, ignoring this and just picking one value or calculating a simple average, this may lead to wrong conclusions.

In my thesis, I followed this approach and used all 10 plausible values for each regression model. This helped me get more reliable results and respected the official OECD guidelines for analyzing PISA data.

Chapter 5: Experimental results

5.1. Technical description of the working environment and libraries used

The analyses were carried out in a Jupyter Notebook environment running on the Anaconda distribution which provide an integrated and intuitive platform for scientific computing in Python. Also, the environment provided a framework for interactive coding, sequential visual analysis and efficient data handling. Thus, for data handling and manipulation, the Pandas library was primarily used, alongside NumPy for numerical operations. The dataset was initially loaded using pyreadstat, which allows efficient reading of SPSS files. Moreover, for visualization, both matplotlib.pyplot and seaborn were employed to rigorous static plot and explore data patterns visually. Furthermore, to build a machine learning pipeline and preprocess the data, scikit-learn components such as Pipeline and tools like shuffle were used. Additionally, the joblib library was used to store trained models efficiently.

5.2. Experiments with tree-based models

This section presents a comparative analysis of the performance of various regression models train on different subsets from PISA 2022. Rather than using the entire dataset, the assessment was focused on data from individual countries to explore how different regression models perform across 9 country-specific subsets used in the analysis. Above all, each subset represents a distinct country with different number of samples and differ data distributions. None of the subsets have a perfectly normal distribution. These variations allow for meaningful comparisons both between subsets and across models in terms of predictive performance.

5.1.1. Assessment 1: Simple Decision Tree Regressor

Hyperparameter	Description	Value
max_depth	The maximum depth of the tree. Limits how deep the tree can grow.	5
ccp_alpha	Complexity parameter used for minimal cost-complexity pruning.	5
criterion	The function to measure the quality of a split.	friendman_mse
min_samples_leaf	The minimum number of samples required to be at a leaf node.	10
min_samples_split	The minimum number of samples required to split an internal node.	2

Table 5.1 Simple Decision Tree Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.54	0.49	0.58	0.53	0.51	0.52	0.63	0.56	0.51
	Read		0.57	0.51	0.58	0.58	0.51	0.50	0.61	0.59	0.52
	Science		0.56	0.52	0.58	0.58	0.47	0.52	0.61	0.57	0.51
VALID	Math		0.45	0.43	0.54	0.45	0.43	0.45	0.60	0.53	0.44
	Read		0.47	0.45	0.52	0.51	0.44	0.42	0.58	0.52	0.45
	Science		0.47	0.45	0.53	0.50	0.41	0.44	0.55	0.52	0.43
TEST	Math		0.47	0.43	0.51	0.44	0.40	0.47	0.60	0.54	0.44

	Read		0.47	0.46	0.50	0.49	0.42	0.47	0.58	0.55	0.45
	Science		0.48	0.46	0.48	0.48	0.39	0.47	0.58	0.54	0.43

Table 5.2 Simple Decision Tree Regressor model results

In the first experiment, a Simple Decision Tree Regressor model was trained, using the hyperparameters values presented in Table 5.1. The model was trained on 9 subsets corresponding to the 9 countries included in the study. The overall performance across these subsets was modest, with an R^2 score of 0.60 for mathematics and 0.58 for reading and science for France. The subsets for Romania and Hungary follow France, where slightly higher R^2 values were recorded. As expected from a simple model, the overall performance was relatively low, because of the limited capacity of decision trees to generalize, especially in a complex and noisy dataset.

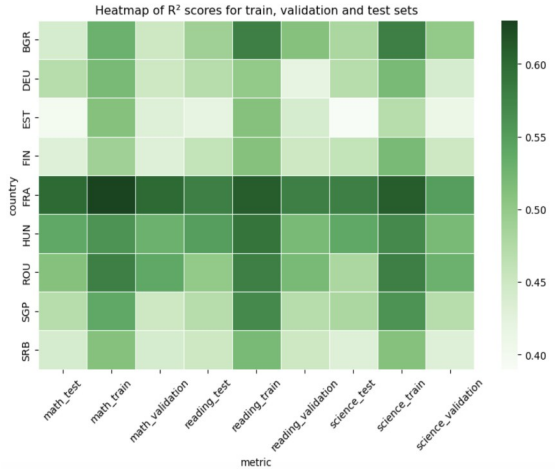


Figure 5.1 Performance heatmap for Simple Decision Tree Regressor model

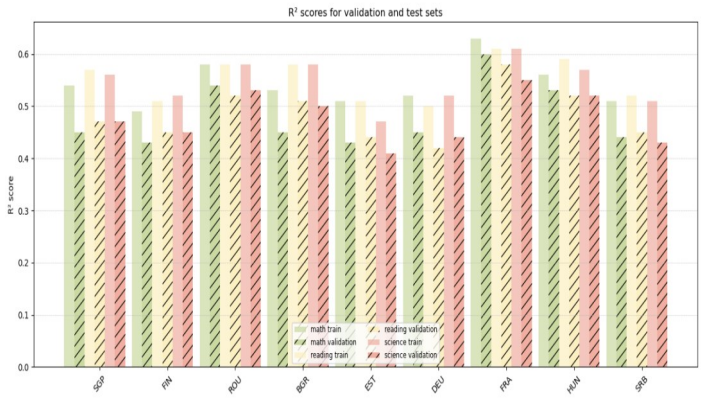


Figure 5.2 R^2 scores by country for training and validation sets – Simple Decision Tree Regressor model

Therefore, the Figure 5.1 displays a heatmap of the R^2 scores achieved by the Simple Decision Tree Regressor model across the training, validation and test sets for the 3 assessed domains: mathematics, reading and science for each of 9 countries included in the study. The R^2 scores has a considerable variability across countries and between subjects. As I previously mentioned, the highest performance was observed in France, particularly in mathematics, where the model reached 0.60 for R^2 score. Moreover, Romania and Hungary also achieved good results across all three domains. In contrast, countries such as Estonia, Serbia and Germany had a weaker performance especially on the validation sets. This fact suggests that the model's ability to generalize is much lower. Also, for most subsets, the R^2 scores on the training data are higher than those on validation and test data. This fact indicates that a tendency toward overfitting. This situation represents a common characteristic of Simple Decision Tree Regressor which are not complex enough to generalize properly in noisy conditions.

5.1.2. Assessment 2: Random Forest Regressor

Hyperparameter	Description	Value
n_estimators	The number of the trees in the forest.	90
max_depth	The maximum depth of each tree. Limits how deep. Each tree can grow.	8
min_saples_leaf	The minimum number of samples required to be at a leaf node.	15

min_samples_split	The minimum number of samples required to split an internal node.	25
max_leaf_nodes	The maximum number of leaf nodes in each tree.	50
max_features	The number of features to consider when looking for the best split.	sqrt

Table 5.3 Random Forest Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.59	0.60	0.62	0.54	0.55	0.55	0.64	0.63	0.53
	Read		0.59	0.63	0.64	0.62	0.55	0.55	0.61	0.64	0.55
	Science		0.59	0.61	0.62	0.59	0.55	0.56	0.61	0.62	0.45
VALID	Math		0.52	0.53	0.56	0.47	0.46	0.50	0.60	0.56	0.44
	Read		0.53	0.56	0.59	0.56	0.46	0.50	0.57	0.56	0.47
	Science		0.52	0.55	0.55	0.53	0.46	0.50	0.56	0.56	0.45
TEST	Math		0.53	0.54	0.56	0.46	0.46	0.52	0.61	0.57	0.46
	Read		0.53	0.57	0.58	0.55	0.44	0.52	0.58	0.57	0.48
	Science		0.53	0.54	0.57	0.52	0.45	0.52	0.57	0.57	0.45

Table 5.4 Random Forest Regressor model results

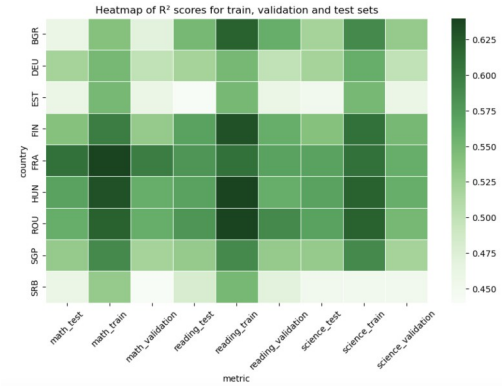


Figure 5.3 Performance heatmap for Random Forest Regressor model

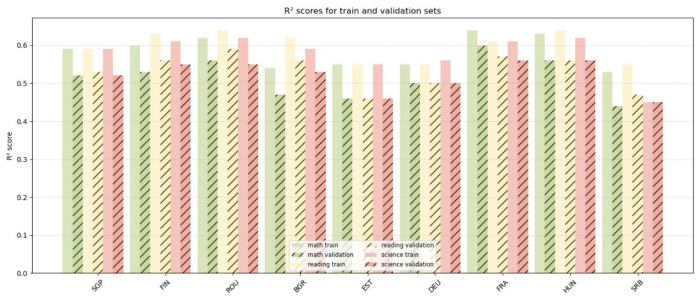


Figure 5.4 R^2 scores by country for training and validation sets – Random Forest Regressor

In this experiment, a Random Forest Regressor model was trained using the hyperparameters values presented in Table 5.3. The model was trained on 9 subsets corresponding to the 9 countries included in the study. Also, the model's performance was evaluated on three data splits: train, validation and test for each three domains: mathematics, reading and science. The results are presented in Table 5.4, Figure 5.3 and Figure 5.4. Compared to the model used in the previous experiment, the Simple Decision Tree Regressor, Random Forest Regressor showed an improvement for R^2 scores, especially on the test sets, where the values are between 0.52 and 0.61. Moreover, the best result was obtained for France, Romania and Hungary, while countries such as Serbia, Estonia and Germany recorded lower scores, but higher than those from the previous model.

5.1.3. Assessment 3: Gradient Boosting Regressor

Hyperparameter	Description	Value
n_estimators	The number of boosting trees in the model.	200
learning_rate	The contribution of each tree to the final prediction.	0.05
max_depth	The maximum depth of the individual regression estimators.	3

min_samples_split	The minimum number of samples required to split an internal node.	20
min_samples_leaf	The minimum number of samples required to be at a leaf node.	10
max_features	The number of features to consider when looking for the best split.	sqrt

Table 5.5 Gradient Boosting Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math		0.76	0.71	0.78	0.68	0.72	0.72	0.79	0.80	0.70
	Read		0.69	0.66	0.73	0.61	0.65	0.68	0.76	0.75	0.62
	Science		0.71	0.68	0.72	0.59	0.64	0.69	0.75	0.76	0.63
VALID	Math	R2	0.75	0.74	0.80	0.76	0.71	0.72	0.77	0.80	0.72
	Read	score	0.69	0.69	0.75	0.70	0.64	0.67	0.74	0.73	0.65
	Science		0.70	0.72	0.74	0.69	0.61	0.70	0.74	0.74	0.66
TEST	Math		0.75	0.72	0.79	0.73	0.72	0.72	0.77	0.79	0.70
	Read		0.68	0.68	0.74	0.67	0.64	0.68	0.72	0.74	0.63
	Science		0.70	0.69	0.71	0.67	0.63	0.70	0.72	0.75	0.63

Table 5.6 Gradient Boosting Regressor model results

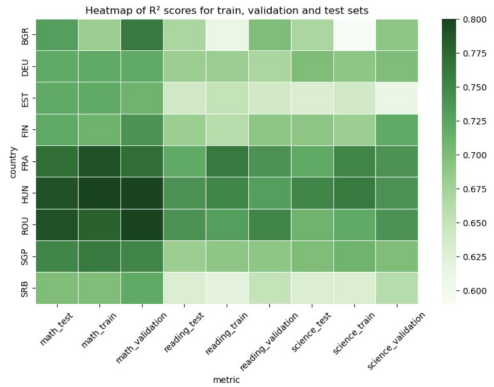


Figure 5.5 Performance heatmap for Gradient Boosting Regressor model

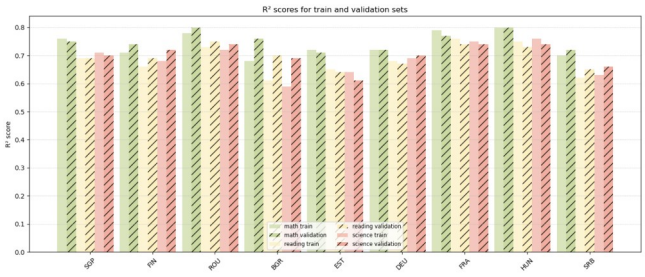


Figure 5.6 R^2 scores by country for training and validation sets – Gradient Boosting Regressor

In this experiment, a Gradient Boosting Regressor model was trained using the hyperparameter values presented in Table 5.5. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.6, Figure 5.5 and Figure 5.6. Gradient Boosting Regressor demonstrated significantly better performance compared to the models trained in previous experiments, achieving high R^2 scores across all datasets. Also. The model performed considerably for mathematics. The R^2 score on test set reached up to 0.79 for Romania and Hungary and 0.77 for France. Thus, the model performed very well in the reading and science domains, where test R^2 scores were consistently around 0.70. For example, on the Romania subset, the R^2 score. Was 0.74 for reading and 0.71 for science, illustrating the idea that Gradient Boosting Regressor can capture complex relationships within the data. Moreover, an important fact is the small gap between training and validation R^2 scores. This may indicate a high capacity for generalization and a reduced risk of overfitting. This observation is displayed in Figure 5.6. Furthermore, this model proved a better performance than the earlier models.

5.1.4. Assessment 4: Light Gradient Boosting Machine Regressor

Hyperparameter	Description	Value
learning_rate	Control the contribution of each tree to the final prediction.	0.01
num_leaves	Maximum number of leaves per tree. Control model complexity.	6
max_depth	Maximum depth of the tree.	3

Table 5.7 Light Gradient Boosting Machine Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.70	0.65	0.72	0.62	0.64	0.65	0.74	0.73	0.63
	Read		0.62	0.61	0.68	0.57	0.58	0.62	0.72	0.70	0.57
	Science		0.65	0.63	0.67	0.56	0.57	0.63	0.72	0.70	0.57
VALID	Math		0.70	0.68	0.74	0.71	0.63	0.66	0.72	0.74	0.65
	Read		0.61	0.64	0.71	0.66	0.57	0.62	0.69	0.68	0.59
	Science		0.65	0.66	0.69	0.65	0.55	0.65	0.70	0.69	0.60
TEST	Math		0.69	0.66	0.72	0.68	0.64	0.66	0.72	0.72	0.63
	Read		0.61	0.63	0.69	0.62	0.58	0.62	0.68	0.68	0.57
	Science		0.64	0.64	0.65	0.63	0.56	0.64	0.68	0.69	0.56

Table 5.8 Light Gradient Boosting Machine Regressor model results

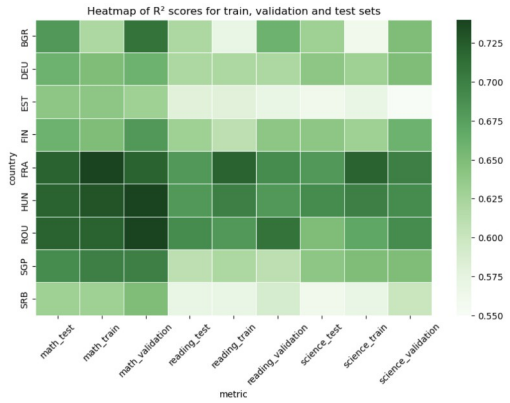


Figure 5.7 Performance heatmap for Light Gradient Boosting Machine Regressor model

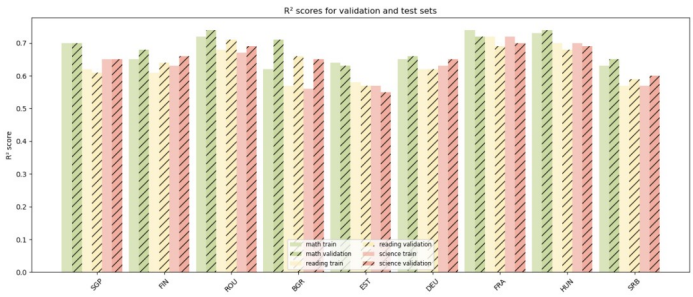


Figure 5.8 R^2 scores by country for training and validation sets – Light Gradient Boosting Machine Regressor

In this experiment, a Light Gradient Boosting Machine Regressor model was trained using the hyperparameter values presented in Table 5.7. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.8, Figure 5.7 and Figure 5.8. The model achieved a weaker performance the gradient boosting model across all three evaluation domains: mathematics, reading and science. The R^2 scores ranged between 0.56 and 0.69 for these three domains, illustrating a low generalization capacity. Also, the highest coefficients of determination were observed in mathematics perditions, where countries such as Romania, France, Hungary and Singapore reached values close to 0.72 on the test set. Similarly, for reading and science predictions, the R^2 score on test set reached up to 0.68 for the same countries. Although LightGBM is known for its efficiency and high performance in complex scenarios, in this case it recorded slightly inferior results compared to the gradient boosting model. A possible explanation might be the higher sensitivity to hyperparameters.

5.1.5. Assessment 5: Extreme Gradient Boosting Regressor

Hyperparameter	Description	Value
n_estimators	The number of boosting trees.	200
max_depth	Maximum depth of each individual tree.	3
learning_rate	Controls the contribution of each tree to the final prediction.	0.04
reg_alpha	L1 regularization term on weights.	0.34
reg_lambda	L2 regularization term on wights.	5

Table 5.9 Extreme Gradient Boosting Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.77	0.74	0.81	0.67	0.69	0.75	0.82	0.83	0.71
	Read		0.76	0.77	0.82	0.75	0.68	0.75	0.80	0.83	0.74
	Science		0.76	0.75	0.81	0.72	0.69	0.76	0.80	0.82	0.72
VALID	Math		0.71	0.67	0.74	0.60	0.61	0.69	0.77	0.76	0.62
	Read		0.69	0.69	0.76	0.69	0.61	0.68	0.75	0.74	0.65
	Science		0.70	0.69	0.75	0.65	0.61	0.68	0.74	0.75	0.63
TEST	Math		0.72	0.69	0.73	0.58	0.60	0.70	0.77	0.77	0.63
	Read		0.71	0.72	0.75	0.67	0.58	0.71	0.75	0.75	0.66
	Science		0.71	0.70	0.72	0.65	0.59	0.71	0.74	0.76	0.62

Table 5.10 Extreme Gradient Boosting Regressor model results

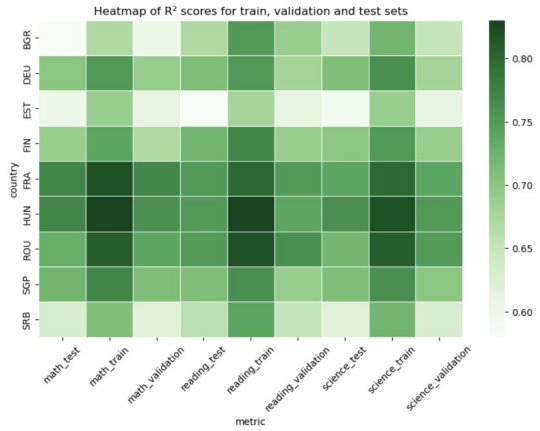


Figure 5.9 Performance heatmap for Extreme Gradient Boosting Regressor model

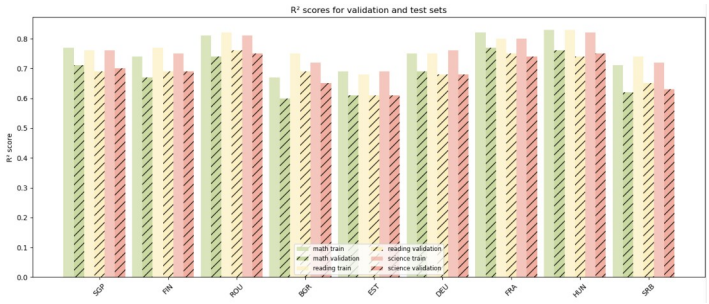


Figure 5.10 R^2 scores by country for training and validation sets – Extreme Gradient Boosting Regressor

In this experiment, an Extreme Gradient Boosting Regressor model was trained using the hyperparameter values presented in Table 5.9. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.9, Figure 5.9 and Figure 5.10. Extreme Gradient Boosting Regressor demonstrated significantly better performance compared to the models trained in previous experiments, achieving high R^2 scores across all datasets. The best results, like the Gradient Boosting Regressor model, were observed in the mathematics domain, where the R^2 score reached 0.77 for France and Hungary and 0.73 for Romania. Overall, R^2 scores for mathematics ranged between 0.60 and 0.77, illustrating the model's ability to capture relationships between input features and students' outcomes. Moreover, in the reading and science

domains, performance remained high with R^2 scores around 0.70 in most country subsets. For instance, in the case of Romania the model achieved 0.75 for reading and 0.72 for science, which represent competitive values with those from the previous experiment. Also, a positive aspect could be the low difference between the training and validation set for R^2 score. This fact suggests high capacity of the model for generalization and low risk of overfitting. The superior performance may be explained by its architecture which is ensemble based.

5.1.6. Assessment 6: Histogram – Based Gradient Boosting Regressor

Hyperparameters	Description	Value
max_iter	The maximum number of boosting trees.	300
learning_rate	Controls the contribution of each tree to the final prediction.	0.01
min_samples_leaf	The minimum number of samples required to be at a leaf node.	100
L2_regularization	Strength of L2 regularization, help prevent overfitting.	10
early_stopping	Stops training when validation performance stops improving.	True
max_depth	Maximum depth of each individual tree.	3

Table 5.11 Histogram Gradient Boosting Regressor hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.62	0.59	0.65	0.55	0.56	0.58	0.59	0.66	0.55
	Read		0.59	0.57	0.61	0.52	0.53	0.54	0.57	0.60	0.51
	Science		0.56	0.55	0.63	0.51	0.49	0.56	0.55	0.62	0.51
VALID	Math		0.61	0.61	0.66	0.63	0.55	0.59	0.61	0.65	0.56
	Read		0.57	0.59	0.64	0.62	0.51	0.53	0.59	0.60	0.53
	Science		0.57	0.58	0.63	0.61	0.50	0.56	0.58	0.62	0.53
TEST	Math		0.61	0.59	0.65	0.60	0.56	0.58	0.59	0.65	0.55
	Read		0.57	0.58	0.61	0.58	0.51	0.55	0.58	0.60	0.51
	Science		0.55	0.57	0.62	0.58	0.50	0.55	0.57	0.60	0.50

Table 5.12 Histogram Gradient Boosting Regressor model results

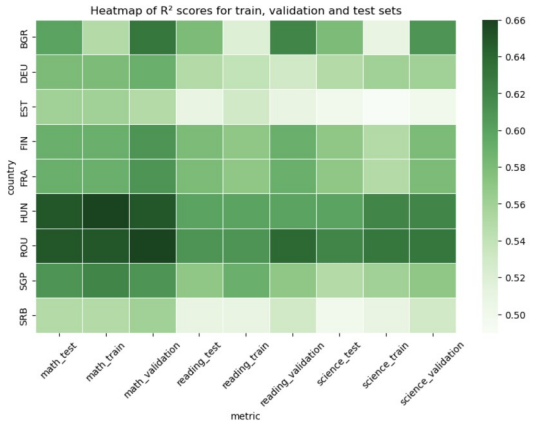


Figure 5.11 Performance heatmap for Histogram Gradient Boosting Regressor model

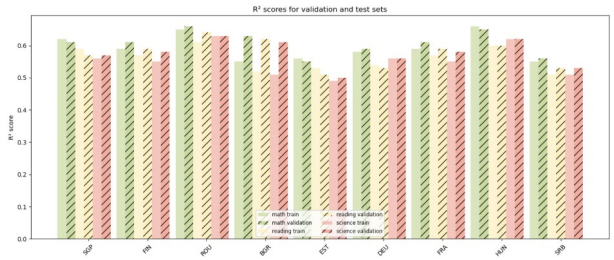


Figure 5.12 R^2 scores by country for training and validation sets – Histogram Gradient Boosting Regressor

In this experiment, a Histogram Gradient Boosting Regression model was trained using the hyperparameter values presented in Table 5.11. The model was also trained on 9 subsets

corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.11, Figure 5.11 and Figure 5.12. Histogram Gradient Boosting Regressor performed weaker the previous model. The R^2 scores on the test set ranged between 0.55 and 0.66. Also, the mathematics domain was once again the best represented. The highest R^2 scores were for Romania and Hungary, 0.66. Furthermore, for the reading and science domains, the scores were relatively close, between 0.55 and 0.63.

5.3. Experiments with linear models

5.3.1. Assessment 7: Linear Regression

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.77	0.74	0.79	0.69	0.76	0.77	0.80	0.83	0.68
	Read		0.77	0.67	0.71	0.53	0.65	0.65	0.72	0.71	0.60
	Science		0.78	0.68	0.72	0.54	0.60	0.68	0.69	0.72	0.60
VALID	Math		0.69	0.76	0.80	0.77	0.75	0.77	0.79	0.84	0.70
	Read		0.69	0.71	0.74	0.66	0.62	0.63	0.69	0.71	0.63
	Science		0.68	0.71	0.72	0.67	0.63	0.65	0.68	0.71	0.61
TEST	Math		0.69	0.75	0.80	0.75	0.76	0.78	0.78	0.83	0.69
	Read		0.69	0.69	0.73	0.60	0.65	0.66	0.69	0.71	0.60
	Science		0.69	0.68	0.74	0.64	0.61	0.67	0.68	0.72	0.60

Table 5.13 Linear Regression model results

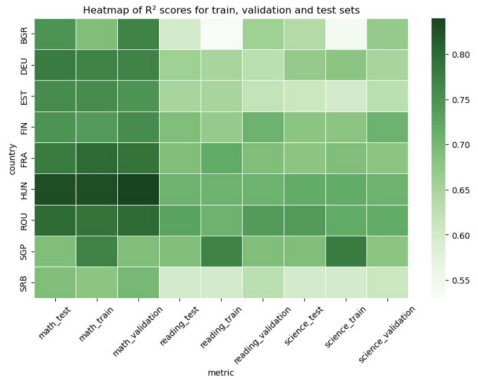


Figure 5.13 Performance heatmap for Linear Regression model

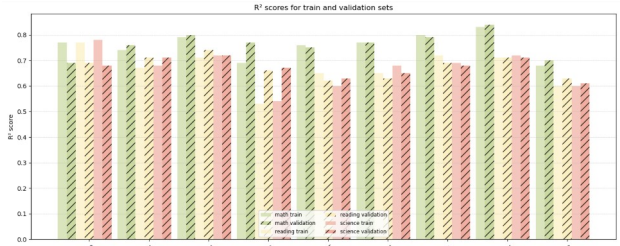


Figure 5.14 Performance heatmap for Linear Regression model

In this experiment, a Linear Regression model was trained. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.13, Figure 5.15 and Figure 5.16. The model demonstrated significantly better performance compared to the models trained in previous experiments, achieving high R^2 scores across all datasets, especially in mathematics domain. The R^2 score on test set reached up to 0.83 for Hungary, 0.80 for Romania and 0.78 for France and Germany.

Also, all subsets achieved scores above 0.69, which indicates the linear relationships between data were captured. For the reading and science domains, the performance is moderate, with test R^2 scores Between 0.60 and 0.74. Romania achieved 0.73 in reading and 0.74 in science, while Hungary achieved 0.71 in reading and 0.72 in science. These results suggest that Linear Regression can capture important information despite its simplicity. A remarkable observation is the difference between training and validation scores, which highlights a good generalization on unseen data and low risk of

overfitting. In conclusion, while the Linear Regression is one of the most basic predictive models, it performed surprisingly well in this context, especially for mathematics.

5.3.2. Assessment 8: Ridge Regression

Hyperparameter	Description	Value
alpha	Regularization strength.	1
max_iter	Maximum number of iterations for the solver to converge.	10000
tol	Tolerance for stopping criteria.	1e-3

Table 5.14 R Ridge Regression model hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.77	0.74	0.80	0.70	0.75	0.76	0.79	0.83	0.69
	Read		0.77	0.76	0.81	0.78	0.75	0.76	0.78	0.83	0.71
	Science		0.78	0.75	0.80	0.75	0.75	0.77	0.78	0.83	0.69
VALID	Math		0.70	0.67	0.71	0.53	0.67	0.67	0.73	0.73	0.61
	Read		0.68	0.69	0.72	0.64	0.67	0.66	0.72	0.73	0.63
	Science		0.70	0.68	0.72	0.60	0.65	0.67	0.70	0.73	0.61
TEST	Math		0.70	0.68	0.71	0.51	0.61	0.68	0.72	0.74	0.59
	Read		0.70	0.70	0.71	0.61	0.61	0.66	0.72	0.74	0.61
	Science		0.70	0.69	0.71	0.60	0.62	0.67	0.70	0.73	0.60

Table 5.15 Ridge Regression model results

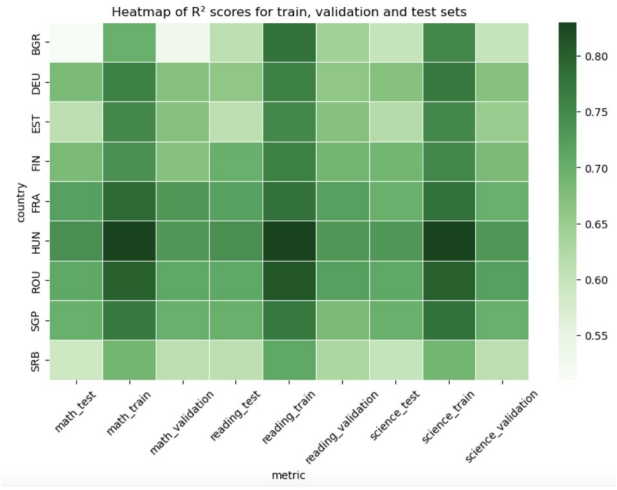


Figure 5.15 Performance heatmap for Ridge Regression model

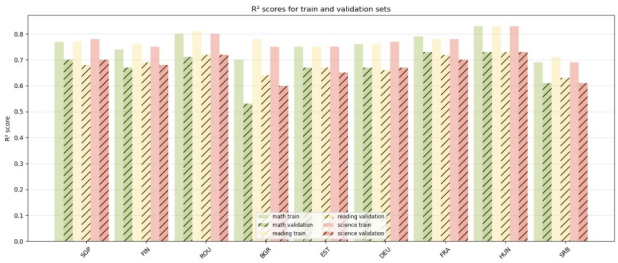


Figure 5.16 R^2 scores by country for training and validation sets – Ridge Regression

In this experiment, a Ridge model was trained using the hyperparameter values presented in Table 5.14. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.14, Figure 5.15 and Figure 5.16. Ridge Regression is a regularized linear model that adds L2 penalty to the loss function to reduce overfitting and improve generalization. The model was assessed on the train, validation and test sets for all three domains: mathematics, reading and science. The model performed well with test R^2 scores ranged between 0.59 and 0.74 across all subjects and countries. Moreover, the highest performance was observed again in the mathematics domain with scores reaching up to 0.74 for Hungary, 0.72 for France and 0.71 for Romania. For reading and science, the model maintained the scores between 0.61 and 0.74. For Romania, the model achieved 0.71 for reading and 0.71 for science, matching the results of more

complex models such as LightGBM and XGBoost. Also, the minor difference between training and validation sets highlights low risk of overfitting, thanks to its L2 regularization.

5.3.3. Assessment 9: Lasso Regression

Hyperparameter	Description	Value
alpha	Regularization strength.	1

Table 5.16 Lasso Regression model hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.74	0.71	0.76	0.65	0.70	0.72	0.76	0.79	0.66
	Read		0.74	0.74	0.78	0.74	0.70	0.71	0.75	0.79	0.68
	Science		0.75	0.72	0.77	0.71	0.71	0.73	0.75	0.79	0.66
VALID	Math		0.70	0.67	0.72	0.60	0.68	0.70	0.75	0.75	0.60
	Read		0.70	0.70	0.74	0.69	0.67	0.68	0.73	0.74	0.62
	Science		0.69	0.69	0.74	0.66	0.66	0.69	0.75	0.74	0.60
TEST	Math		0.71	0.69	0.72	0.57	0.65	0.71	0.73	0.77	0.60
	Read		0.71	0.71	0.73	0.66	0.65	0.70	0.73	0.76	0.62
	Science		0.72	0.70	0.72	0.66	0.64	0.71	0.72	0.75	0.60

Table 5.17 Lasso Regression model results

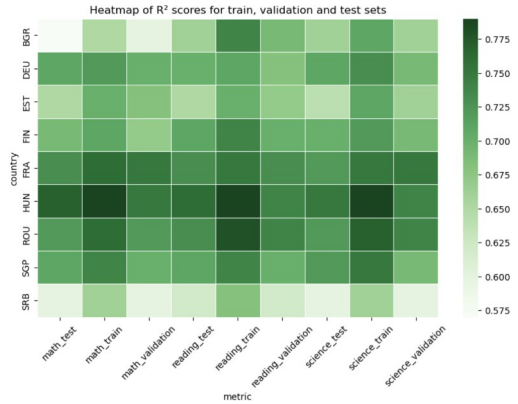


Figure 5.17 Performance heatmap for Lasso Regression model

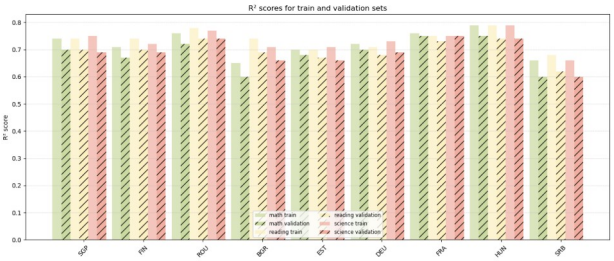


Figure 5.18 R^2 scores by country for training and validation sets – Lasso Regression

In this experiment, a Lasso Regression model was trained using the hyperparameter values presented in Table 5.16. The model was also trained on 9 subsets corresponding to the 9 countries included in the study. The results obtained are illustrated in Table 5.16, Figure 5.17 and Figure 5.18. Lasso Regression is a regularized linear model that adds L1 regularization to prevent overfitting and for feature selection, where reduce some coefficients to zero. The model was assessed on the train, validation and test sets for all three domains: mathematics, reading and science. The model performed well with test R^2 scores ranged between 0.57 and 0.77 across all subjects and countries. Moreover, the highest performance was observed again in the mathematics domain with scores reaching up to 0.77 for Hungary, 0.73 for France and 0.72 for Romania. For reading and science, the model maintained the scores between 0.60 and 0.76. Romania and Hungary stood out with scores around 0.73. What's more the results could be compared with those achieved by Ridge Regression, suggesting that Lasso Regression works well on this dataset. Another strength of Lasso is reflected

in the balance between training and validation set, which highlights a good generalization on unseen data and a low risk of overfitting.

5.3.4. Assessment 10: Elastic Net Regression

Hyperparameter	Description	Value
alpha	Regularization strength.	1
L1_ratio	The balance between L1 and L2 regularization.	0.5

Table 5.18 Elastic Net Regression model hyperparameters

Set	Domain		SGP	FIN	ROU	BGR	EST	DEU	FRA	HUN	SRB
TRAIN	Math	R2 score	0.74	0.71	0.76	0.65	0.70	0.72	0.76	0.79	0.65
	Read		0.74	0.73	0.78	0.74	0.70	0.71	0.75	0.79	0.67
	Science		0.74	0.72	0.77	0.71	0.71	0.73	0.74	0.79	0.66
VALID	Math		0.70	0.67	0.72	0.60	0.68	0.70	0.75	0.75	0.59
	Read		0.70	0.70	0.74	0.69	0.67	0.68	0.73	0.74	0.62
	Science		0.69	0.69	0.74	0.66	0.67	0.70	0.72	0.75	0.60
TEST	Math		0.72	0.69	0.72	0.57	0.65	0.72	0.74	0.77	0.60
	Read		0.71	0.71	0.73	0.67	0.63	0.71	0.74	0.76	0.62
	Science		0.72	0.70	0.73	0.66	0.64	0.71	0.72	0.75	0.60

Table 5.19 Elastic Net Regression model results

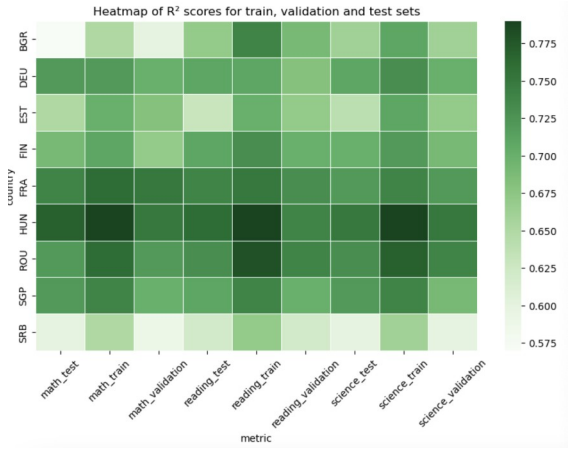


Figure 5.19 Performance heatmap for Elastic Net Regression model

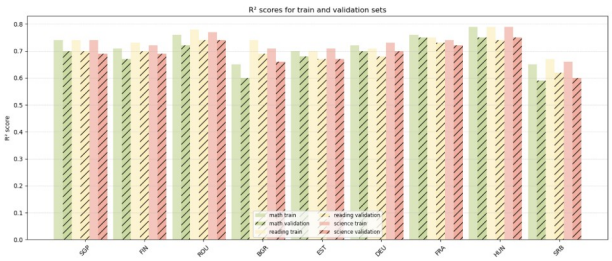


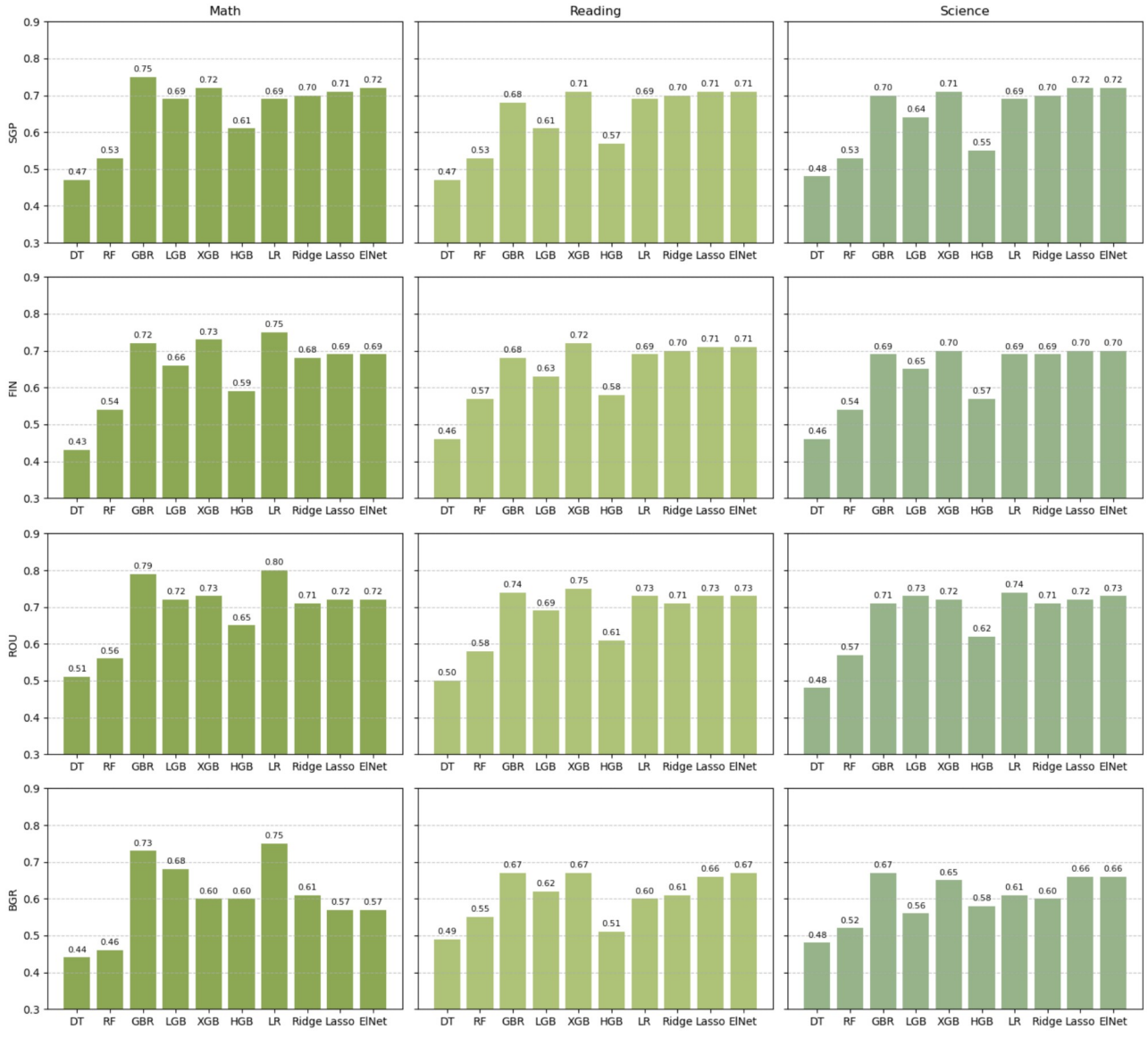
Figure 5.20 R^2 scores by country for training and validation sets – Elastic Net Regression

In this experiment, Elastic Net Regression model was trained using the hyperparameters values presented in Table 5.18. The model was trained on 9 subsets corresponding to the 9 countries included in the study. Also, the model's performance was evaluated on three data splits: train, validation and test for each three domains: mathematics, reading and science. The results are presented in Table 5.19, Figure 5.19 and Figure 5.20. Elastic Net Regression is a linear model that combines both L1 and L2 regularization, keeping the model simple and capturing complex patterns. The model performed well across all domains, particularly with strong results in mathematics, where R^2 scores reached 0.77 for Hungary, 0.74 for France and 0.72 for Romania. Also, in reading and science domains, the model showed good results. For instance, R^2 scores in reading ranged between 0.62 and 0.76 and for science between 0.60 and 0.75 across various country subsets. Moreover, an important strength of ElastinNet is the close values between training and validation scores, illustrating the model's ability to generalize well without overfitting.

Chapter 6: Analysis of model performance across countries

This chapter presents a detailed analysis of how the regression models performed across the selected countries included in the PISA 2022 study. Each country's dataset was treated as a separate subset, allowing for a clear comparison of model performance in different educational contexts. For each country and domain (mathematics, reading, and science) ten regression models were trained and evaluated using the R^2 score. The results are shown in the graphs below, illustrating how model performance varied from one country to another.

Furtermore, the chapter explores the importance of individual features in predicting student scores. For every subset and domain, the top 10 most relevant features are presented. Also the frequency of each feature appeared across models in Table 6.1. Therefore, this chapter highlight key factors that influence student performance. Additionally, cosine similarity analysis was applied to compare the structure of feature importance between countries. This provides a general perspective on which countries share similar factors and where major differences exist.



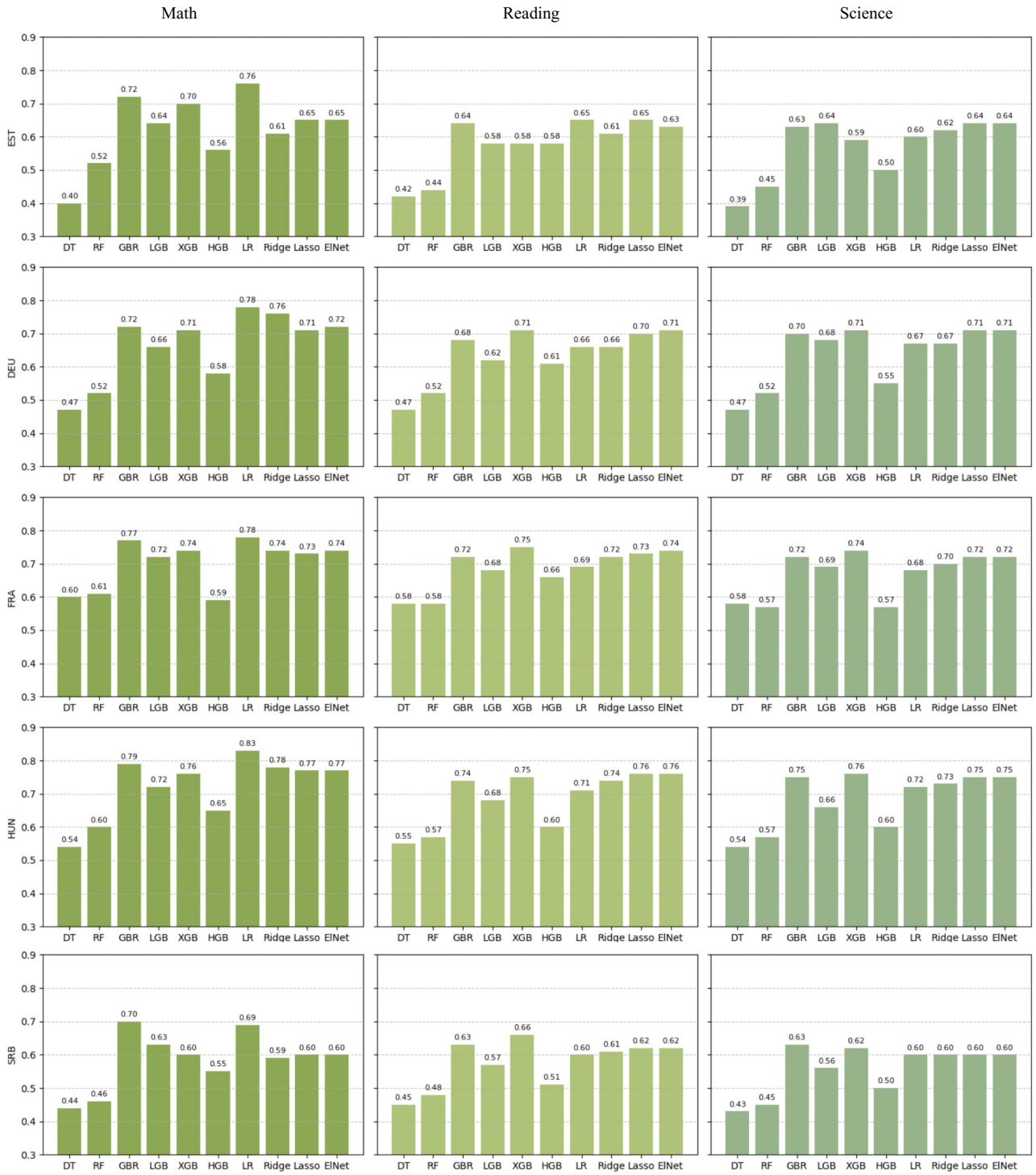


Figure 6.1 Comparison of R^2 scores across models

As a part of the experimental phase, the ten selected regression models were individually trained and evaluated on the data subsets corresponding to the nine countries included in the PISA2022 dataset. Wach subset was treated as an independent dataset, representing student responses for a specific country. The objective was to identify the best models for a national context across the three main PISA domains such as mathematics, reading and science.

The models tested include ensemble models (Simple Decision Tree, Random Forest, Gradient Boosting, Light Gradient Boosting Machine, Extreme Gradient Boosting , Histogram Based Gradient Boosting) and linear models (Linear Regression, Ridge, Lasso, ElasticNet) highlighting a balanced comparison between complex and simple approaches. Furthermore, the performance of the model

was evaluated using R^2 score on the test set, as a measure for each model. As a result, the models showed a significant variability in model performance across countries. The most performant model were the ensemble models such as Gradient Boosting Regressor and Extreme Gradient Boosting Regressor where both regression models are based on relatively the same underlaying algorithm. Also, these two models, consistently performed within the range associated with the best results, between 0.70 and 0.80, for most countries. For instance, R^2 scores reached up to 0.79 for Hungary and France and 0.77 for Romania. On the other hand, linear models also performed surprisingly well. R^2 scores for Romania, Hungary and France were again among the highest. For example, Linear Regression achieved 0.80 for Romania, 0.83 for Hungary followed by other countries with scores above 0.70 for mathematics domain. These vales suggest a strong linear relationship between input and output values. In contrast, countries as Serbia, Bulgaria and Estonia recorded lower scores, particularly in reading and science, illustrating highest data variability and features with low predictive power. These differences between sets, highlight the need for specific models for each country.

Country	Math (test R^2 score)	Reading (test R^2 score)	Science (test R^2 score)
SGP	**GBR – 0.75** XGB - 0.72 Ridge – 0.70 Lasso – 0.71 ElNet – 0.72	**XGB – 0.71** Ridge – 0.70 **Lasso – 0.71** **ElNet – 0.71**	GBR – 0.70 XGB – 0.71 Ridge – 0.70 **Lasso – 0.72** **ElNet – 0.72**
FIN	GBR – 0.72 XGB - 0.73 **LR – 0.75**	**XGB – 0.72** Ridge – 0.70 Lasso – 0.71 ElNet – 0.71	XGB – 0.70 Lasso – 0.70 **ElNet – 0.70**
ROU	**GBR – 0.79** LGB – 0.72 XGB – 0.73 **LR – 0.80** Ridge – 0.71 Lasso – 0.72 ElNet – 0.72	GBR – 0.74 **XGB – 0.75** LR – 0.73 Ridge – 0.71 Lasso – 0.73 ElNet – 0.73	GBR – 0.71 LGB – 0.73 XGB – 0.72 **LR – 0.74** Ridge – 0.71 Lasso – 0.72 ElNet – 0.73
BGR	GBR – 0.73 **LR – 0.75**	R2_score under 0.70	R2_score under 0.70
EST	GBR – 0.72 XGB – 0.70 **LR – 0.76**	R2_score under 0.70	R2_score under 0.70
DEU	GBR – 0.72 XGB – 0.71 **LR – 0.78** **Ridge – 0.76** Lasso – 0.71 ElNet – 0.72	**XGB – 0.71** Lasso – 0.70 **ElNet – 0.71**	GBR – 0.70 XGB – 0.71 Lasso – 0.71 ElNet – 0.71

FRA	**GBR – 0.77** LGB – 0.72 XGB – 0.74 **LR – 0.78** Ridge – 0.74 Lasso – 0.73 ElNet – 0.74	GBR – 0.72 **XGB – 0.75** Ridge – 0.72 Lasso – 0.73 ElNet – 0.74	GBR – 0.72 **XGB – 0.74** Ridge – 0.70 Lasso – 0.72 ElNet – 0.72
HUN	**GBR – 0.79** LGB – 0.72 XGB – 0.76 **LR – 0.83** **Ridge – 0.78** Lasso – 0.77 ElNet – 0.77	GBR – 0.74 XGB – 0.75 LR – 0.71 **Ridge – 0.74** **Lasso – 0.76** **ElNet – 0.76**	**GBR – 0.75** **XGB – 0.76** LR – 0.72 Ridge – 0.73 **Lasso – 0.75** **ElNet – 0.75**
SRB	GBR – 0.70	R2_score under 0.70	R2_score under 0.70

Table 6.1 Best model per country and subject (test R^2 score)

After training of 10 regression models, their performance was measured using the R2 score. The score was different across countries and domains. Also, for feature importance analysis, 6 models were kept, 3 tree-based (Gradient Boosting Regressor, Extreme Gradient Boosting Regressor and LightGBM) and 3 linear models (Ridge, Lasso and ElasticNet Regression) ensuring a balance comparison between ensemble and linear models. Furthermore, from each model the most important features were extracted using permutation importance. This method is used to assess the impact of each feature for the prediction. This algorithm shuffles the values of a single feature and measure how much, the performance increase or decrease. Thus, if the performance decreases significantly, the feature is important. If the performance is almost unchanged, the feature has a low influence. So, this method is useful to capture the real contribution of each feature for the entire subset. Furthermore, for each of these models, the most important features were selected for each domain using permutation importance method. In addition, to figure out which features are important in each subset, there were kept only those that appeared in at least 3 out of the 6 models. This idea, helped to reduce the noise and focus on the ones that were regularly important. Then, for each of those features, the average importance was calculated.

The process was repeated for each country and domain and the top 10 most important features for each were kept. Figure 6.2 to 6.10 present the most important features for each subset and each domain.

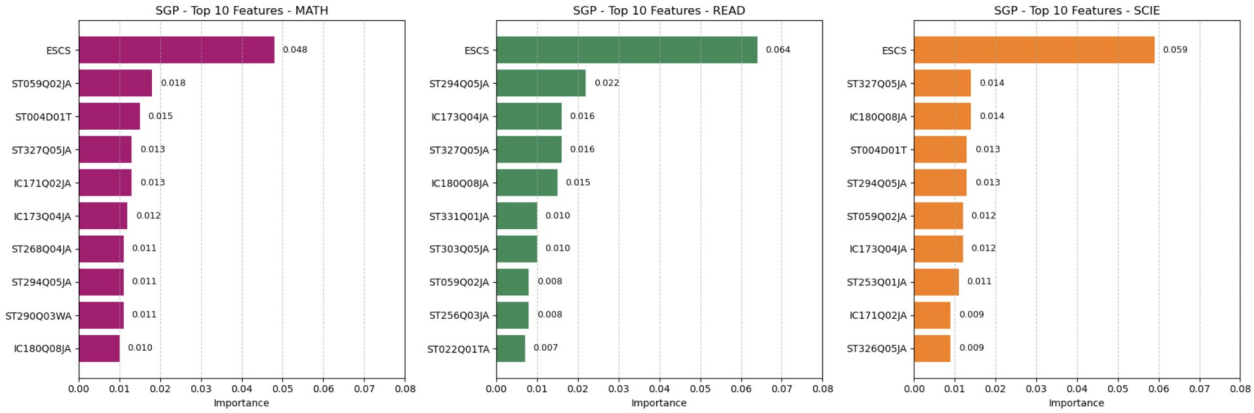


Figure 6.2 Top 10 key features for Singapore by domain

For instance, the most important feature for Singapore is ESCS (Index of economic, social and cultural status) showing the higher score in mathematics and science, while the next has a score of 0.014. Thus, this strong contrast suggests that economic, social and cultural status has a dominant influence for Singapore. In contrast, the rest of the features has almost similar scores, which indicates, they provide important information, but these features do not have same impact as ESCS. ST294Q05JA appears in all three domains and highlights how often students exercise before school, suggesting a possible connection between academic performance and exercise. ST327Q05JA is about students expectations for future education levels, may suggest that those who set higher academic goals could perform better. What's more, IC180Q08JA illustrates how often students share false information on social media without flagging it. This feature may be an indicator for critical thinking or digital responsibility. Moreover, ST059Q02JA represents the total number periods per week may reflect the main exposure to learning.

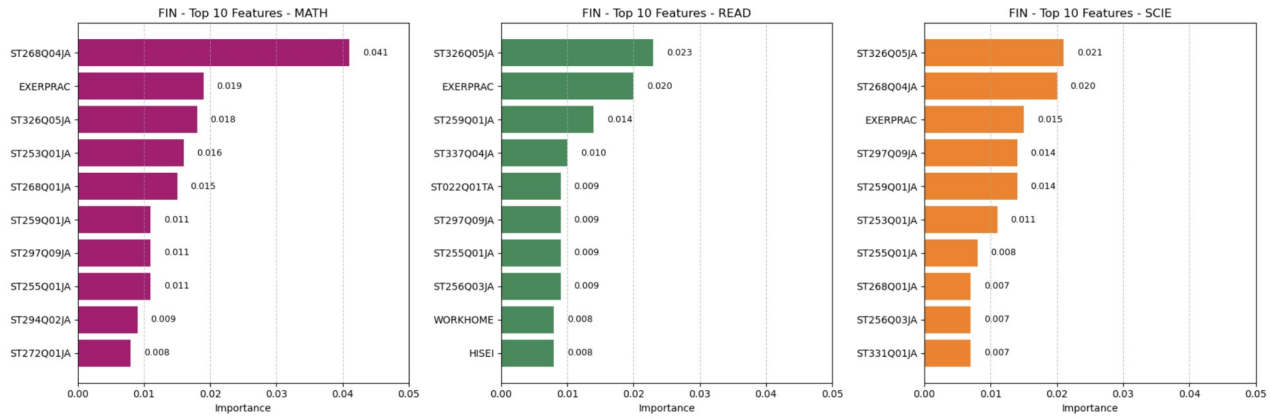


Figure 6.3 Top 10 key features for Finland by domain

Furthermore, for the Finland subset, we observed distinct characteristics compared to those identified for Singapore. Additionally, the key features are different across the domains. In mathematics, the most influential feature is ST268Q04JA ("Mathematics is easy for me."), highlighting the importance of self-perception of their abilities in this field. Also, EXERPRAC ("Exercise or practice a sport before or after school.") and ST326Q05JA ("This school year, how many hours/day use [digital resources] for: For leisure before and after school.") illustrates that physical exercises and daily routine may play an important role in academic performance. Moreover, for reading domain EXERPRAC, ST326Q05JA and ST259Q01JA have the first three places. These suggest that physical activity, the digital tools and family status may shape the reading skills. Also, for science domains the most important features were EXERPRAC, ST268Q04JA, ST253Q01JA and ST259Q01JA.

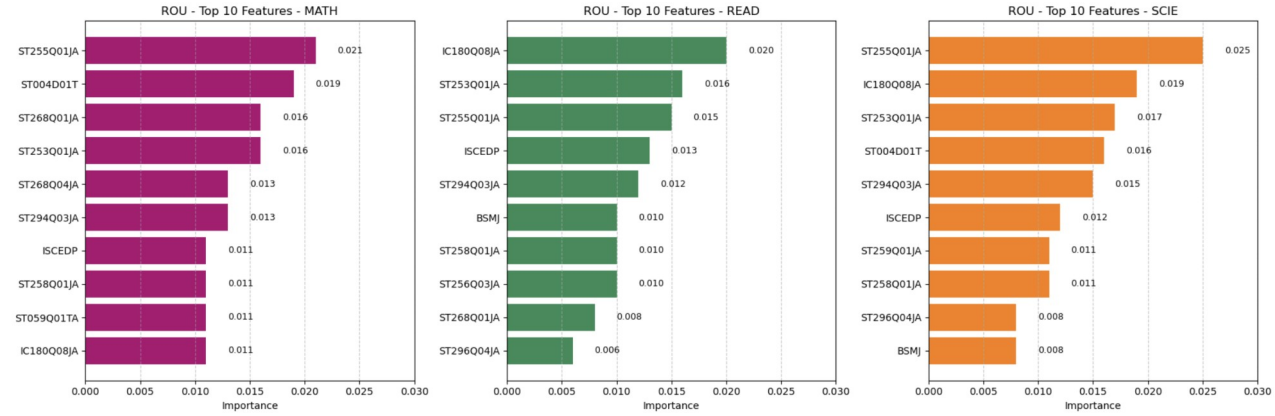


Figure 6.4 Top 10 key features for Romania by domain

For the Romania subset, the top 10 features identified for each domain contribute to the model predictions. For instance, in the mathematics domain appears ST255Q01JA ("How many books are there in your [home]?"), ST004D01T ("Student (Standardized) Gender"), ST268Q01JA ("Mathematics is one of my favourite subjects.") and ST268Q04JA ("Mathematics is easy for me."). These features highlight the role of educational resources at home and self confidence in mathematics which shape the academic performance. In contrast, the variables such as ST294Q03JA ("How many days/wk before school: Work in the household or take care of family members.") and ST258Q01JA ("In the past 30 days, how often did you not eat because there was not enough money to buy food?") illustrate the negative factors in the academic development of Romanian students. These facts show that family responsibilities and food insecurity may negatively affect the performance. Moreover, the feature IC180Q08JA ("I share made-up information on social networks without flagging its inaccuracy") across all domains, may draw a relationship between digital activity and academic success. What's more, the variable ISCEDP (parents' level of education) is frequently observed. This fact suggests the influence of family education background. Also in reading, features like ST256Q03JA (number of contemporary books at home) and BSMJ (expected occupational status) are displayed. These features bring to light the connection between cultural environment and reading performance. In science, in addition to the previous features mentioned, there are ST259Q01JA (family social standing) and ST296Q04JA (time spent on homework), showing a mix between personal effort and environment. In conclusion, the subset for Romania reveals that educational, social and psychological factors may influence the student performance. Also, the results explained the need for improvements in education, considering students living conditions.

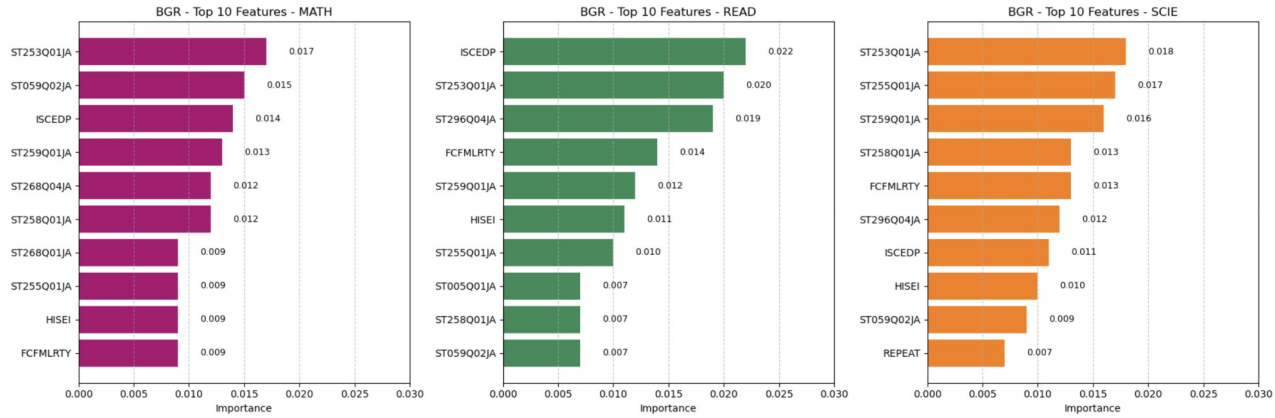


Figure 6.5 Top 10 key features for Bulgaria by domain

For Bulgaria, the feature ST253Q01JA ("How many digital devices with screens are there in your home?") stands out as the most influential across all academic domains. This underlines how important digital access at home is when it comes to students academic success. That finding is alligned with two other indicators: ST255Q01JA (the number of books at home) and ST259Q01JA (how students perceive their family's social standing). Therefore, the home environment and socioeconomic conditions play an important role in shaping educational outcomes. Also,in mathematics, features like ST059Q02JA (total number of weekly class periods) and ST268Q04JA (agreement that math is easy) highlight the combined impact of classroom exposure and student confidence. Meanwhile, the appearance of FCFMLRTY (familiarity with finance concepts) and HISEI (highest parental occupation status) suggests that both practical knowledge and family background are closely linked to performance in this subject. For reading, the most influential feature is ISCEDP (parents' education level), pointing to the strong effect of parental education. Other relevant factors include ST296Q04JA (time spent on homework) and FCFMLRTY once again, showing that both good habits and exposure to real-world concepts matter across subjects. The inclusion of ST005Q01JA (mother's education level) provides that maternal influence plays also an important role. Familiar features such as ST253Q01JA, ST255Q01JA and ST259Q01JA reappear, joined by REPEAT (grade repetition). The consistent presence of variables like HISEI, FCFMLRTY, and ISCEDP across domains underlines the steady role of both economic status and parental education. In conclusion, Bulgaria's profile paints a clear picture: family background, digital access and student attitudes consistently shape academic results. The fact that these factors show up again and again across different subjects makes them strong candidates for targeted policy and educational support.

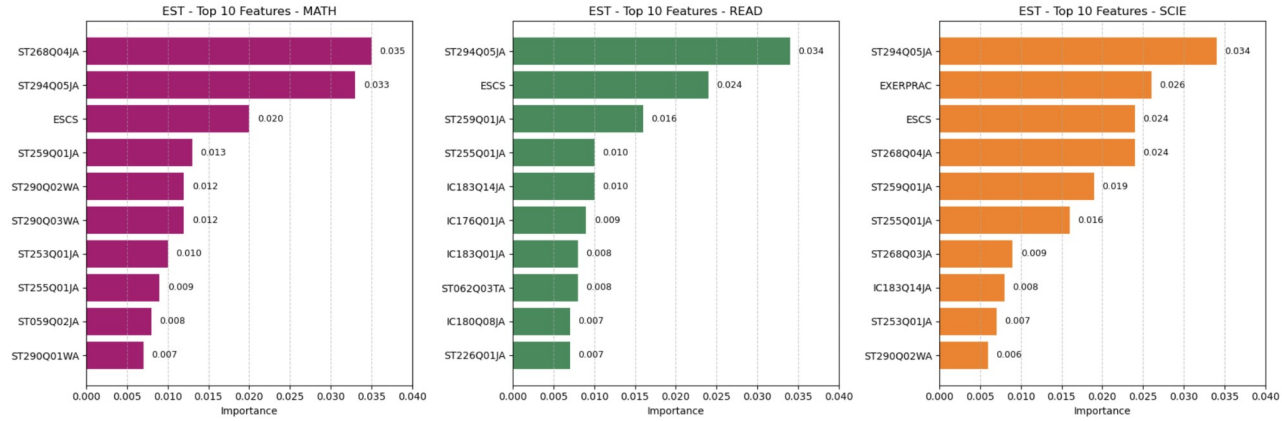


Figure 6.6 Top 10 key features for Estonia by domain

In Estonia, the main focus is on student independence, healthy habits, and digital skills. In math. As we cn see, the strongest feature is ST268Q04JA ("Mathematics is easy for me"), showing that how students see their own abilities may have an impact on how well they do. This is followed by ST294Q05JA (doing physical activity before school), and some about confidence in solving everyday math problems such as ST290Q03WA, ST290Q02WA and ST290Q01WA. Also, features like ESCS and ST259Q01JA (how students see their family's social status) also play a role. In reading, top features like IC183Q14JA, IC176Q01JA, and IC183Q01JA are all linked to digital literacy. Additional, features about arriving late to school (ST062Q03TA) and how long a student has been at the same school (ST226Q01JA) suggest that routines and stability may influence reading results. Again, background factors like ESCS and the number of books at home continue to show up.

Furthermore, in science field, physical activity is a strong predictor. ST294Q05JA and EXERPRAC show a link between staying active and doing well in school. In addition, interest in the subject also matters. ST268Q03JA ("Science is one of my favourite subjects") shows that liking science may increase the performance. Also, digital skills show up here too, with IC183Q14JA (creating a program in Scratch, Python, etc.) connected to better outcomes.
To sum it up, Estonia represents a mixture of features like self-confidence, physical activity and family background.

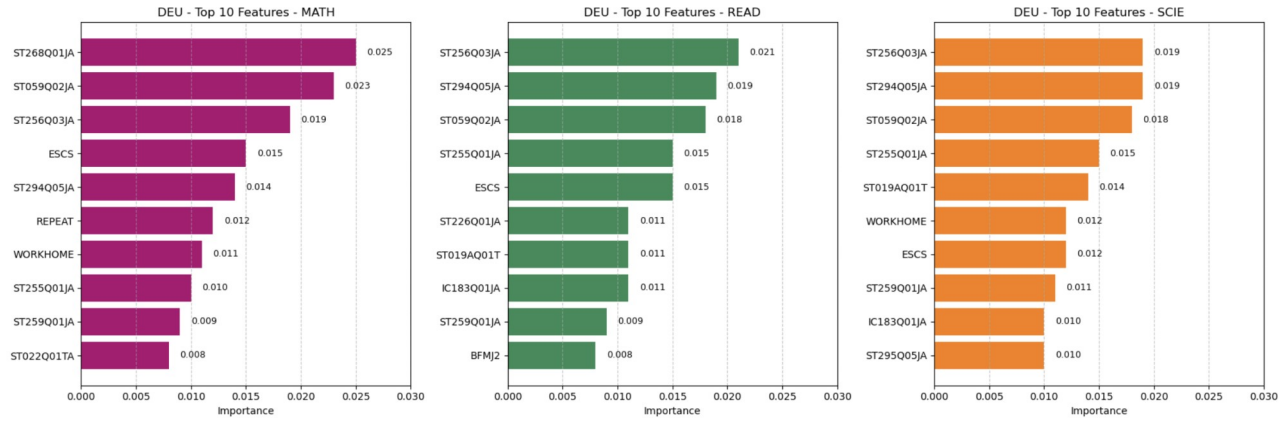


Figure 6.7 Top 10 key features for Germany by domain

In Germany, the top predictive features reveal a link between students habits, their family background, and also their personal interest in learning. For instance, in mathematics domain, one of the most influential indicators is ST268Q01JA ("Mathematics is my favourite subjects"), which show the student's attitude for this subject. This attitude may shape their performance. Moreover, other key factors include ST059Q02JA (number of weekly class periods) and ST256Q03JA (number of contemporary literature books at home), suggesting that both structured school exposure and access to educational resources at home play important roles. Additional features such as ESCS, ST294Q05JA (physical activity before school), and REPEAT (grade repetition) illustrate the influence of personal routines on academic outcomes. Furthermore, in reading, features like ST256Q03JA, ST294Q05JA and ST059Q02JA show up again, painting the idea that home learning environments and physical well-being matter. What's more, the feature IC183Q01JA (searching for information online), pointing to the importance of tech skills in Germany's education system. Meanwhile, BFMJ2 (father's occupational status) and ST019AQ01T (student and parents' country of birth) suggest that social and cultural background are meaningful factors in shaping educational field.
In addition, the science field follows a similar pattern with features like ST294Q05JA, ST256Q03JA, and ST059Q02JA. These features show the value of consistent academic structure and home habits. Also, WORKHOME (involvement in chores or caregiving) and IC183Q01JA reappear here, revealing the relevance of responsibility and digital competence across learning domains.
Overall, Germany shows a balanced mix of thinking skills, personal habits and family background. Moreover, features like ST256Q03JA, ESCS and ST294Q05JA suggest that student performance depends not just on school, but also on how students live, what they enjoy, and the support they get at home.

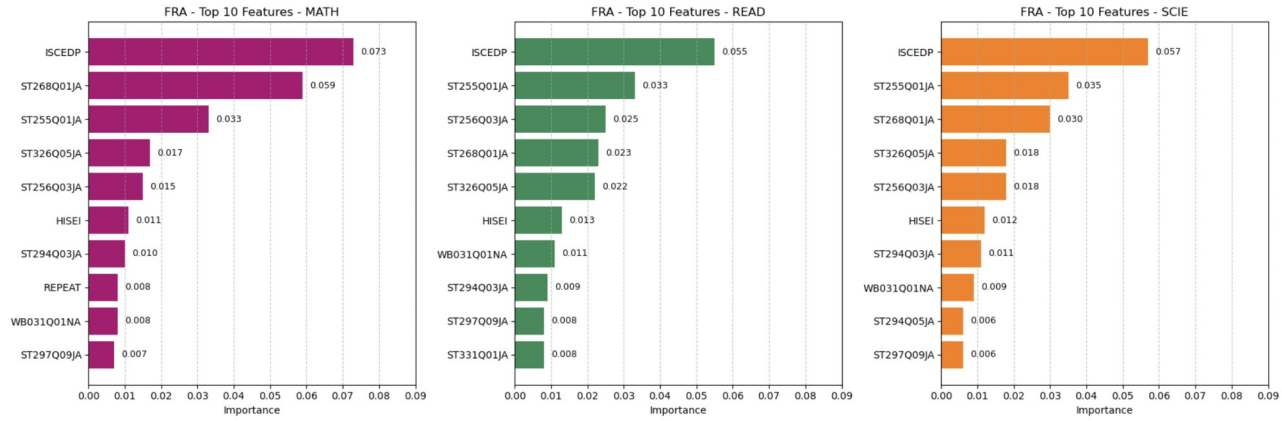


Figure 6.8 Top 10 key features for France by domain

For France, in math domain, the most important feature is ISCEDP, which shows how much education the students' parents have. This suggests that parental education plays an important role in how well students do. Next is ST268Q01JA, which reflects whether students enjoy math. Thus, enjoying the subject may help performance. Also, ST255Q01JA (number of books at home) and ST256Q03JA (modern literature at home) show that a rich learning environment at home also matters. Whats's more physical activity (WB031Q01NA) and grade repetition (REPEAT) are also linked to performance. This fact reveals the importance of well-being and academic background. In addition, for the next domain, reading, many of the same features show up again. Features such as ISCEDP, ST255Q01JA, and ST256Q03JA highlight that the home environment is important. It's interesting that ST268Q01JA (about liking math) also appears here, showing that motivation and positive attitudes help across subjects. More features such as how much effort students put into the test (ST331Q01JA) and whether they take extra math lessons (ST297Q09JA), illustrate that personal effort and studying support make a difference too in academic performance.

Next domain, science, are very similar. Features like ISCEDP, ST255Q01JA and ST268Q01JA are again the top factors. Lifestyle and digital habits also play a role. Also, features like ST326Q05JA (using technology for leisure) and physical activity show that how students spend their time outside of class may model the academic performance.

In summary, France has a clear and consistent academic profile. Across all subjects, we see that student success is influenced by family background, home resources, personal habits and motivation. suggests that success in school depends on a blend of cognitive, emotional, and environmental factors.

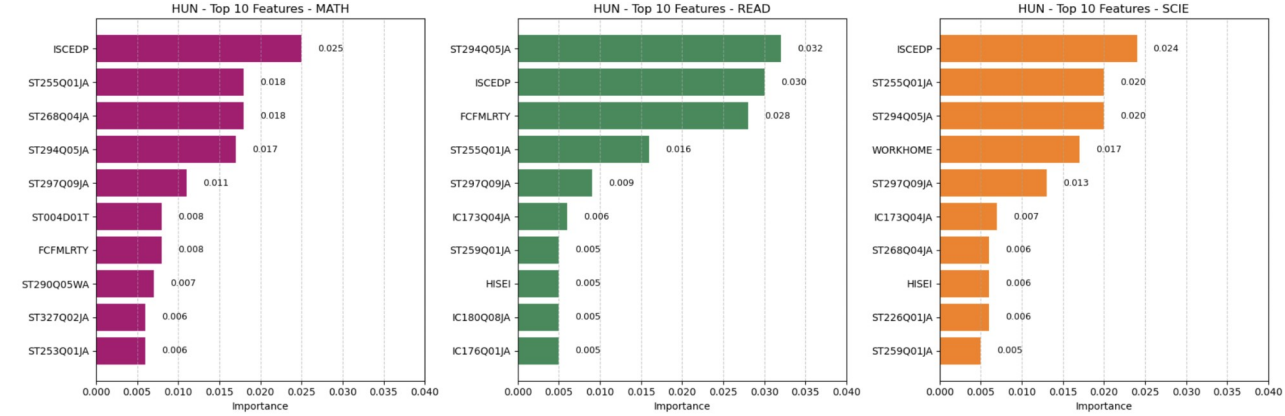


Figure 6.9 Top 10 key features for Hungary by domain

In Hungary, even though the top predictors change depending on the subject, a clear pattern shows up. The features combines things like internal motivation, family support, and how ready students are to use digital tools.

For math, one of the strongest signs of success is whether students feel confident—like saying “Math is easy for me”(ST268Q04JA).Moreover, other important feature is ST255Q01JA ("How many books students have at home?") and ST294Q05JA (physical activity before school), pointing to both thinking skills and lifestyle. Also, ST297Q09JA (additional math classes) appeared. This fact means that there are some students who might need more support in mathematics.

When it comes to reading, physical activity (ST294Q05JA) stands out again. But this time is the most important factor. It might help with focus and concentration. Moreover, other key influences include financial literacy (FCFMLRTY), how educated the parents are (ISCEDP) and how students use digital tools such as spotting fake information online (IC180Q08JA) or using tech in class (IC173Q04JA). These suggest that background knowledge, critical thinking and responsible tech use all matter.

Therefore, science follows a similar pattern. Physical activity and books at home still play an important role. Another interesting factor is how much students help at home (WORKHOME), which might affect how focused they are at school. Also, IC173Q04JA (digital use in class), ST226Q01JA (how long students have been enrolled) and ST297Q09JA (extra support like tutoring) are highlighted in science.

In summary, Hungary has a mix of strengths like books and family educationand digital tools.

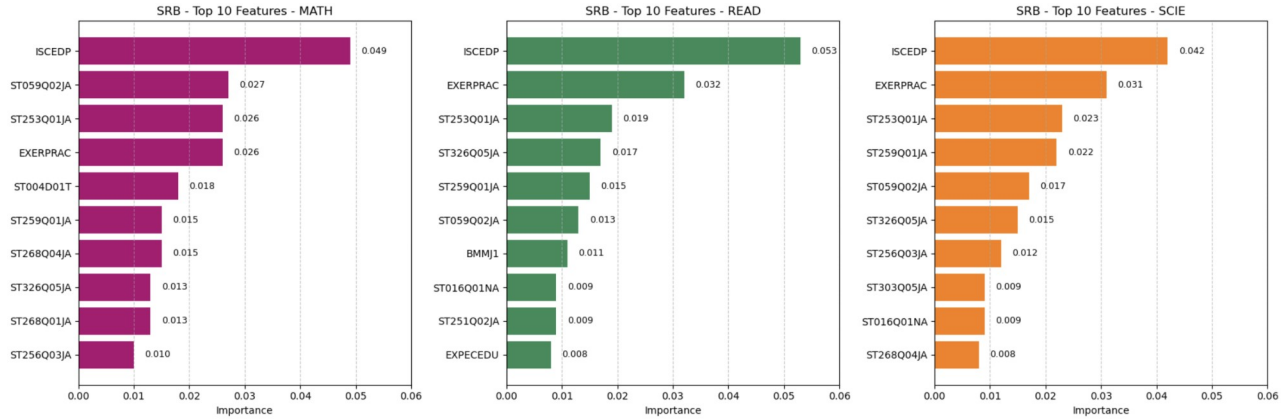


Figure 6.10 Top 10 key features for Sebia by domain

In Serbia, the most relevant predictors across math, reading, and science consistently highlight the influence of educational structure and students' environments.

In mathematics, ISCEDP (parental education level) leads the list, followed closely by ST059Q02JA (weekly class periods), suggesting that both family background and instructional time significantly affect learning. Other top features like EXERPRAC (physical activity), ST253Q01JA (number of digital devices at home) and ST259Q01JA (student's perceived family status) point to the role of everyday context and resources. This pattern holds in reading as well. ISCEDP remains a top predictor, showing how important educational background is. Also, ST251Q02JA (mopeds or motorcycles at home) and EXPECEDU (students' expected education level) suggest that material conditions and personal aspirations may influence interest in reading.

In science, many of the same variables show up: ISCEDP, EXERPRAC and ST253Q01J, indicating a reliable group of key factors. Additional, features like ST016Q01NA (life satisfaction) and ST303Q05JA (rigidity of opinions) introduce a psychological dimension, showing that emotional and attitudinal traits also play a part in learning outcomes.

In summary, Serbia's academic profile is shaped by a blend of formal factors (school time, parental education) and personal context (tech access, lifestyle, well-being). The repeated appearance of similar features across subjects suggests a reliable pattern in how learning is shaped in the country.

Feature	MATH	READ	SCIE	TOTAL	Question
ST255Q01JA	7	7	7	21	How many books are there in your [home]?
ST259Q01JA	5	6	7	18	Now think about where you would place your family on this scale. Where would you say your family stands at this time?
ISCEDP	5	5	5	15	Levels of education programmes (ISCED 2011)
ST253Q01JA	6	3	6	15	How many [digital devices] with screens are there in your [home]?
ST059Q02JA	5	4	4	13	Total number of [class periods] per week for all subjects, including mathematics
ST294Q05JA	4	4	5	13	How many days/wk before school: Exercise or practise a sport (e.g. running, cycling, aerobics, soccer, skating, [country-specific])
ST256Q03JA	3	5	4	12	How many of these books at [home]: Contemporary literature
ST268Q04JA	7	0	4	11	Agree/disagree: Mathematics is easy for me.
ST268Q01JA	6	2	2	10	Agree/disagree: Mathematics is one of my favourite subjects.
ST326Q05JA	3	3	4	10	This school year, how many hours/day use [digital resources] for: For leisure before and after school
ESCS	3	3	3	9	Index of economic, social and cultural status
HISEI	2	4	3	9	Highest parental occupational status (ISEI) based on 4-digit human coded ISCO
ST297Q09JA	3	3	3	9	[Additional math instruction] received: I do not participate in [additional mathematics instruction]
IC180Q08JA	2	4	2	8	Agree/disagree: I share made-up information on social networks without flagging its inaccuracy.
EXERPRAC	2	2	3	7	Exercise or practice a sport before or after school
ST258Q01JA	2	2	2	6	In the past 30 days, how often did you not eat because there was not enough money to buy food?

ST294Q03JA	2	2	2	6	How many days/wk before school: Work in the household or take care of family members
ST004D01T	4	0	2	6	Student (Standardized) Gender
IC173Q04JA	1	2	2	5	How often use [digital resources] in lessons in: [Computer science], [information technology], [informatics] or similar lessons.
FCFMLRTY	2	2	1	5	Familiarity with concepts of finance
WORKHOME	1	1	2	4	Working in household/take care of family members before or after school
ST296Q04JA	0	2	2	4	How much time spent on homework in: Total time for all homework in all subjects, including subjects not listed above
ST022Q01TA	1	2	0	3	What language do you speak at home most of the time?
REPEAT	2	0	1	3	Grade repetition
WB031Q01NA	1	1	1	3	This school year, on average, on how many days do you attend physical education classes each week?
ST226Q01JA	0	2	1	3	How long have you been enrolled at this school?
IC183Q01JA	0	2	1	3	Can do with [digital resources]: Search for and find relevant information online
ST327Q05JA	1	1	1	3	Which of the following qualifications do you expect to complete: [ISCED level 5]
ST331Q01JA	0	2	1	3	How much effort did you put into doing well on the PISA test?
ST016Q01NA	0	1	1	2	Overall, how satisfied are you with your life as a whole these days?
ST019AQ01T	0	1	1	2	In what country were you and your parents born? You
IC176Q01JA	0	2	0	2	How often used [digital resources] to: See my grades or results from specific assignments (e.g. homework or tests)
IC183Q14JA	0	1	1	2	Can do with [digital resources]: Create a computer program (e.g., in [Scratch®], [Python®], [Java®])
ST290Q02WA	1	0	1	2	How confident in math tasks: Calculating how much more expensive a computer would be after adding tax
ST303Q05JA	0	1	1	2	Agree/disagree: I think there is only one correct position in a disagreement.
IC171Q02JA	1	0	1	2	How often use out of school: Smartphone (i.e. mobile phone with Internet access)
ST290Q03WA	2	0	0	2	How confident in math tasks: Calculating how many square metres of tiles you need to cover a floor
BSMJ	0	1	1	2	Expected occupation status (free response)- 4 digits

ST005Q01JA	0	1	0	1	What is the [highest level of schooling] completed by your mother?
ST295Q05JA	0	0	1	1	How many days/wk after school: Exercise or practise a sport (e.g. running, cycling, aerobics, soccer, skating, [country-specific])
ST251Q02JA	0	1	0	1	How many of these items are there at your [home]: Mopeds or motorcycles
BMMJ1	0	1	0	1	Mother's occupational status (ISEI) based on 4-digit human coded ISCO
ST327Q02JA	1	0	0	1	Which of the following qualifications do you expect to complete: [ISCED level 3.3]
ST290Q05WA	1	0	0	1	How confident in math tasks: Solving an equation like 6x[sup]2[/sup]+5=29
ST294Q02JA	1	0	0	1	How many days/wk before school: Study for school or homework
BFMJ2	0	1	0	1	Father's occupational status (ISEI) based on 4-digit human coded ISCO
ST268Q03JA	0	0	1	1	Agree/disagree: [Science] is one of my favourite subjects.
ST272Q01JA	1	0	0	1	On 1-10 scale, rate quality of mathematics instruction this school year? Quality of mathematics instruction?
ST337Q04JA	0	1	0	1	In your school, how often participate in: Debate [club]
ST062Q03TA	0	1	0	1	In the last two full weeks of school, how often: I arrived late for school
ST059Q01TA	1	0	0	1	Number of [class periods] per week in mathematics
ST290Q01WA	1	0	0	1	How confident in math tasks: Working out from a [train timetable] how long it would take to get from one place to another
EXPECEDU	0	1	0	1	Highest expected educational level

Table 6.2 Feature descriptions and their frequency [1]

Table 6.2 illustrates the description of the most important features identified in the regression models applied to the 9 national subsets extracted from the PISA 2022 dataset. The columns MATH, READ and SCIE indicate the number of instances of each feature in the top 10 for each domain, while la TOTAL column represents the cumulative frequency across all three domains. As a result, ST255Q01JA ("How many books are there in your [home]?") is distinguished by its consistent appearance in almost all 9 subsets and across all three domains. It was recorded 21 times. This fact suggests that the home environment may be a strong predictor of academic performance. Another key feature is ST259Q01JA ("Now think about where you would place your family on this scale. Where would you say your family stands at this time?") and ISCEDP ("Levels of education programmes (ISCED 2011)") highlighting a significant importance between family background and student achievement. Also, another important feature is ST253Q01JA ("How many [digital devices] with screens are there in your [home]?") with 15 appearances across all domains. This feature illustrates

the role of technology at home plays a strong role. Moreover, ST253Q01JA appears more in math and science which means being comfortable with digital tools could help students to understand better logical problems or technical subjects. What's more, ST059Q02JA which measures the total number of class periods per week, appeared 13 times. This situation shows that the time spent at school still plays an important role in education and in predicting student performance. Thus, it suggests a potential link between academic exposure and students' performance. Moreover, another key feature is ST294Q05JA, pointing to a strong importance of physical activity before school. It may suggest that students who exercise regularly tend to have a better focus and better academic results. Also, the feature ST256Q03JA highlights the role of reading environments into academic success. Additionally, the variables ST268Q04JA ("Agree/disagree: Mathematics is easy for me.") and ST268Q01JA ("Agree/disagree: Mathematics is one of my favourite subjects.") with 11 and 10 appearances suggest the idea of self-perception and interest in mathematics. These features are important predictors in math domain, confirming that motivation and confidence can play an important role in the learning process.

In the figures below, Figure 6.11 and Figure 6.12 illustrate the importance of features for each country-domain combination. Thus, the average importance of each feature for each subset can be visualized, based on the previously calculated values. In the first heatmap all important features are displayed. Each cell represents the average importance of that variable for a combination country-domain. As can be seen, darker colors indicate higher importance, while lighter colors correspond to lower influence values. Due to the large number of features, a comprehensive representation is displayed in Figure 6.12, providing a better overview. This heatmap shows only the most frequent features across countries and domains. For instance, the feature ISCEDP (parental education level) frequently appears with high importance in predicting scores, especially for reading and science domains in countries such as Estonia, France and Singapore. Also, each square in the heatmap shows the mean importance of a particular variable for a given country-domain pair, with darker colors indicating greater influence. For instance, features such as ISCEDP (parental education level) and ST255Q01JA (number of books at home) stand out with higher importance in several countries, especially in the reading and science domains. This supports existing research showing that family background and access to learning resources are strong predictors of academic outcomes.

Because the complete heatmap in Figure 6.11 contains a very large number of variables, Figure 6.12 was created as a filtered version, highlighting only the most frequently important features across all countries and domains. This more compact representation makes it easier to observe key trends and interpret the results more clearly. The same variables that appear most often with high importance in Figure 6.11 such as ISCEDP, number of books or digital devices at home and perceptions about school or family are the ones shown in Figure 6.12.

These visualizations reinforce the idea that variables related to students' socio-economic background, attitudes toward learning, and home resources are among the most consistent and influential predictors across countries. At the same time, the importance of each feature can still vary depending on the national and educational context.

In summary, the heatmaps are an essential tool in this analysis. They not only show which features are most relevant for prediction, but also allow for cross-national comparisons and a deeper understanding of what influences student achievement globally.

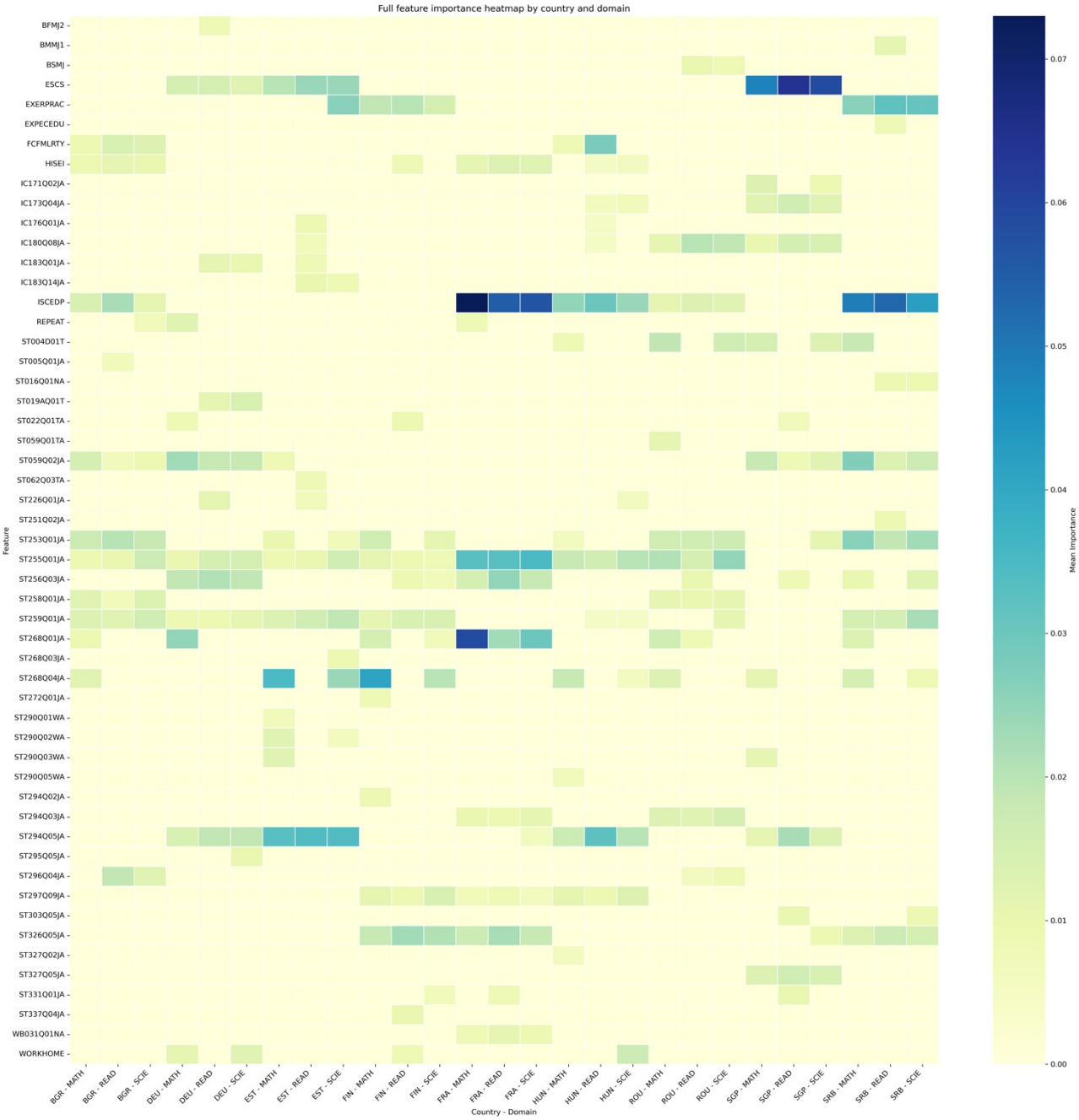


Figure 6.11 Heatmap of importance features by country and domain

Figure 6.12 Heatmap of the top 8 most frequent features by country and domain

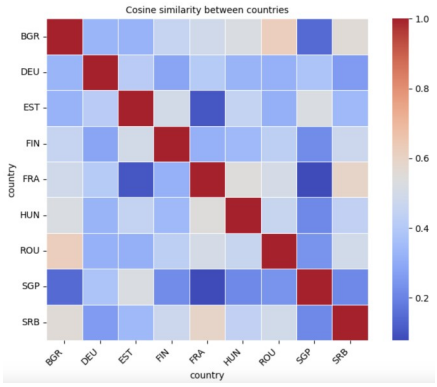


Figure 6.13 Cosine similarity between countries

Country	BGR	DEU	EST	FIN	FRA	HUN	ROU	SGP	SRB
BGR	1.00	0.33	0.32	0.45	0.49	**0.52**	**0.62**	0.15	**0.56**
DEU	0.33	1.00	0.41	0.28	0.41	0.32	0.32	0.38	0.26
EST	0.32	0.41	1.00	0.50	0.10	0.45	0.32	**0.53**	0.34
FIN	0.45	0.28	**0.50**	1.00	0.32	0.34	0.43	0.22	0.48
FRA	0.49	0.41	0.10	0.32	1.00	**0.54**	**0.51**	0.08	0.60
HUN	**0.52**	0.32	0.45	0.34	**0.54**	1.00	0.46	0.21	0.44
ROU	**0.62**	0.32	0.32	0.43	0.51	0.46	1.00	0.24	0.44
SGP	0.15	0.38	**0.53**	0.22	0.08	0.21	0.24	1.00	0.21
SRB	**0.56**	0.26	0.34	0.48	0.60	0.44	0.49	0.21	1.00

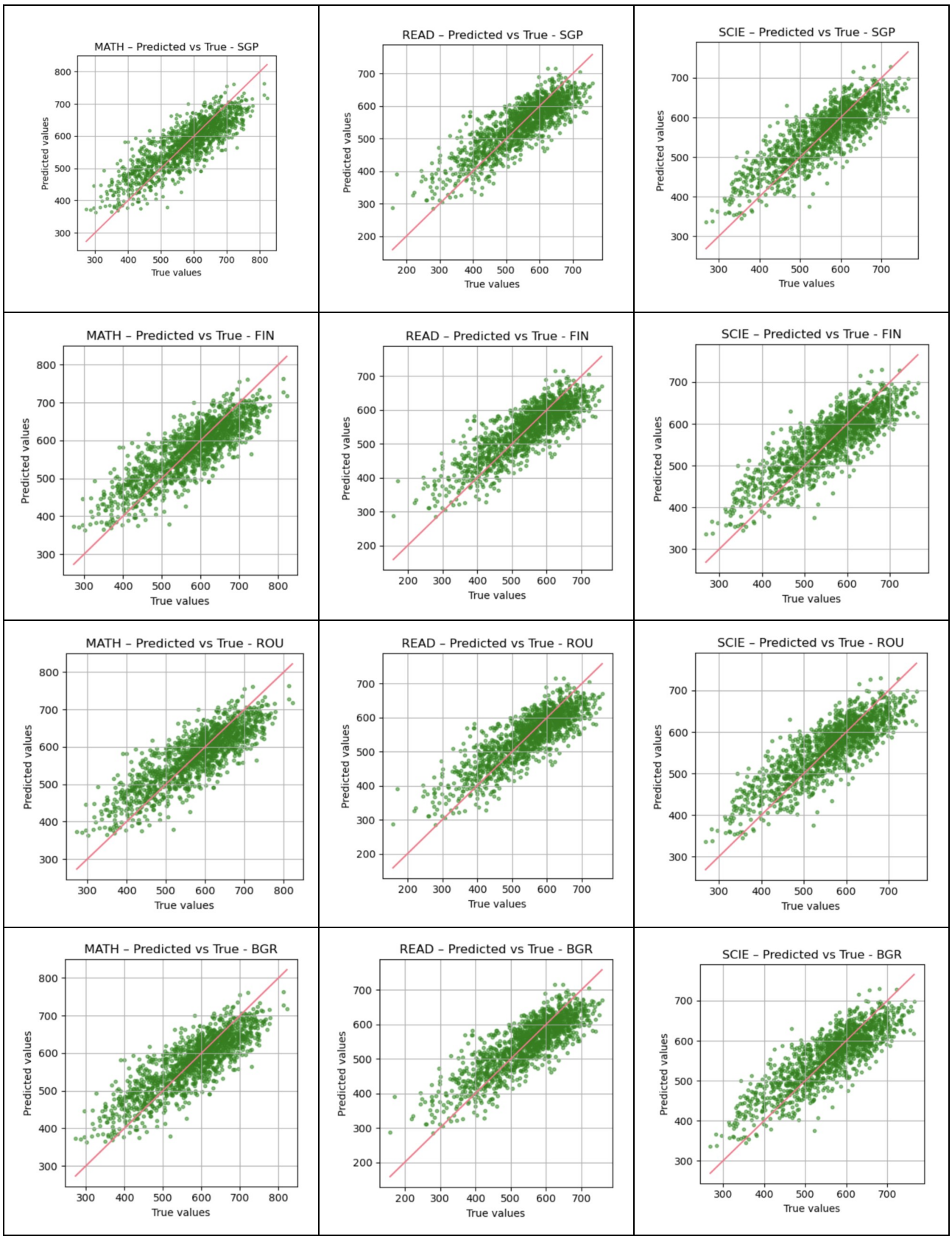
Table 6.3 Cosine similarity

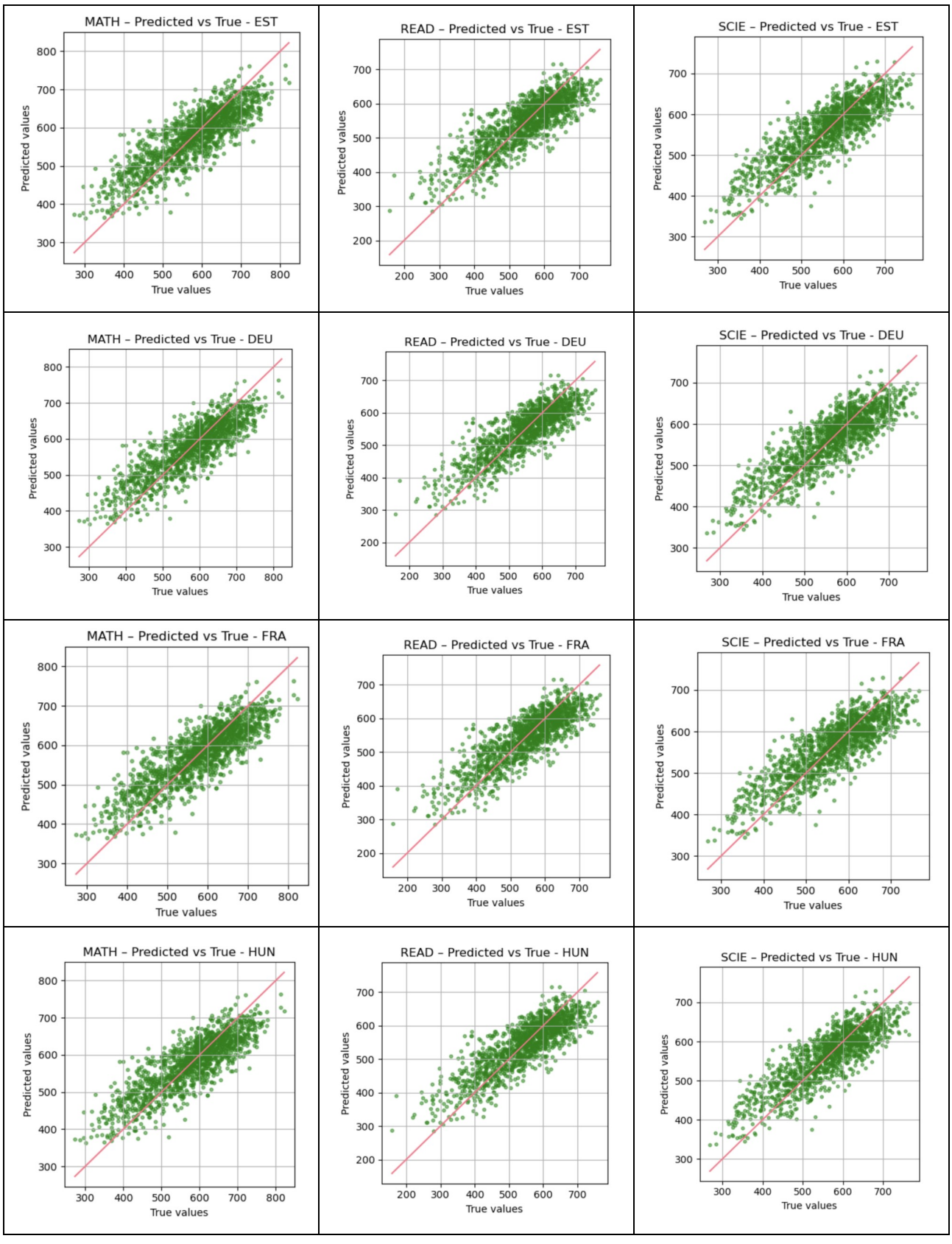
Cosine similarity represents a common method used to measure how similar two non-zero vectors are in a space with many dimensions. Also, it works especially when the direction of the vectors matters more than their size. This fact highlights the fact that it is useful for comparing things like future importance values. Thus, cosine similarity looks at the angle between two vectors. Also, if the point is in the same direction, the similarity is high, otherwise is low.

This gives a value between -1 and 1. In this research, cosine similarity was used to compare feature importance values for each country, based on regression models. Cosine similarity method is not the only fast to compute but works well on high dimensional data [62].

In this study, cosine similarity was used to compare the average feature importance vectors obtained from regression models trains on PISA 2022 subsets for each country. Thus, the countries with similar predictive structures, based on which features are considered the most important.

The heatmap from Figure 5.33 represents a visual representation of how similar the studied countries are in terms of their feature importance, based on the regression models trained for PISA 2022 data. Therefore, a strong similarity between two countries suggests that student and input factors had a similar impact on performance across subjects such as mathematics, reading and science. Also, from the Figure 5.33, we can observe that Romania, Hungary and France show relatively high similarity. This is consistent with the results from earlier experiments, where the same models performed well in these countries. This fact suggests that educational outcomes in these contexts may be influenced by similar features. For instance, the highest similarity is observed between Romania and Bulgaria with a score of 0.62, highlighting a strong similarity in the predictive factors for both countries. Similarly, France and Serbia show a high similarity, 0.60, while France and Hungary follow closely with 0.54. Therefore, these high values illustrate that these countries likely share common educational patterns. In contrast, France and Singapore show the lowest similarity score, 0.08, revealing the expectations that Singapore system is a high performing educational system. Also, Singapore shows weak similarity with the most European countries, including Romania (0.24) and Hungary (0.21), demonstrating its distinct profile. Curiously, although Estonia has medium similarity with Singapore (0.53) and Finland (0.50), it is opposite with France (0.10).


MATH – Predicted vs True - SGP
READ – Predicted vs True - SGP
SCIE – Predicted vs True - SGP
MATH – Predicted vs True - FIN
READ – Predicted vs True - FIN
SCIE – Predicted vs True - FIN
MATH – Predicted vs True - ROU
READ – Predicted vs True - ROU
SCIE – Predicted vs True - ROU
MATH – Predicted vs True - BGR
READ – Predicted vs True - BGR
SCIE – Predicted vs True - BGR
Predicted values
True values


MATH – Predicted vs True - EST
READ – Predicted vs True - EST
SCIE – Predicted vs True - EST
MATH – Predicted vs True - DEU
READ – Predicted vs True - DEU
SCIE – Predicted vs True - DEU
MATH – Predicted vs True - FRA
READ – Predicted vs True - FRA
SCIE – Predicted vs True - FRA
MATH – Predicted vs True - HUN
READ – Predicted vs True - HUN
SCIE – Predicted vs True - HUN
Predicted values
True values

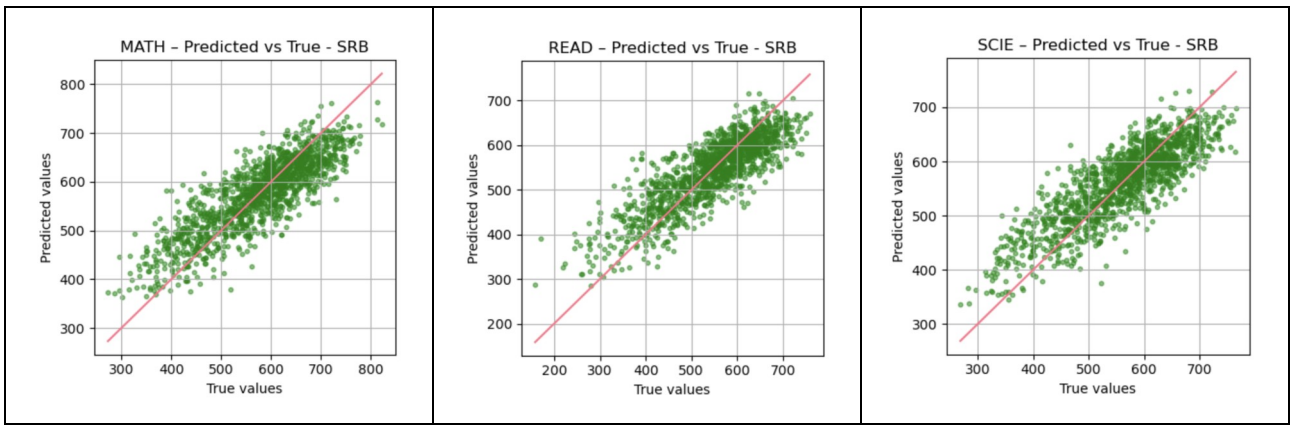


Figure 6.14 Predicted vs true values with ElasticNet

The Figure 6.14 shows the predicted values versus the true values for each country and domain. Most of the data points for each of the three domains cluster around the identity line or red diagonal, which stands for ideal prediction, indicating a general increasing tendency. This Figure 6.14 shows that the model is capturing the fundamental patterns in the data. Moreover, we can see the dispersion of points also highlights that the model does not always produce a perfect prediction. This fact is highlighted in the cases where the predicted scores deviate above or below the line. In mathematics, there is a correlation between the expected and actual scores. It is seen the distribution around the diagonal. Also, the diagonal point concentration is somewhat higher in the reading domain than in mathematics. Also, comparing across domains, it appears that the reading domain often shows tighter clustering along the diagonal, suggesting better model fit and possibly stronger predictive relationships between the features and the reading scores. This might imply that reading performance is influenced by more consistent or well-captured factors, such as home environment, access to books, which are among the key features used in the model. In contrast, mathematics predictions show slightly more variability. The science domain tends to fall somewhere in between the two in terms of accuracy. For instance, Singapore shows similar results in all domains, while the plots for Romania and Bulgaria display more scattered points, suggesting that model performance may vary by country. Therefore, these differences may reflect differences in educational systems, data quality or cultural influences that may model the predictability of student performance.

In general, the points are closely grouped along the line, showing that the model performed well. Germany and France also display good alignment between predicted and true values, although with a bit more spred in the mathematics domain. Hungary shows a similar pattern. The science plot appears slightly more dispersed, suggesting that the features used might not fully capture the variation in science performance for Serbian students.

Overall, these figures highlight that while the model performs consistently across various countries and domains. Factors such as national education systems, cultural influences and sample size likely play a role in this variability. Moreover, these visualizations reinforce the importance of evaluating model performance separately for each context, rather than assuming uniform behavior across all datasets.

Chapter 7 : Conclusions and Future work

7.1 Conclusions

This paper aimed to perform predictive modeling of students' academic performance using the PISA 2022 dataset. The focus was on estimating scores in mathematics, reading, and science for nine countries, as well as extracting the top 10 most important features for each country and domain. The subsets used in the analysis were those for the following countries: Singapore, Finland, Romania, Bulgaria, Estonia, Germany, France, Hungary, and Serbia. Furthermore, the analysis was carried out by training and evaluating ten regression models for each national subset. The overall objective was to identify the most efficient predictive models, as well as the features with the greatest influence on students' results.

Thus, after training the 10 models for each subset, weaker performance was observed for models such as simple decision tree, random forest, linear regression, and histogram gradient boosting. The models were evaluated using the R^2 score, and the best-performing ones were selected for feature importance analysis. Therefore, after training 6 tree-based models and 4 linear regression models, the decision was made to exclude from the feature importance analysis those models that performed poorly, continuing with only 6 models. These 6 models generally performed well on all analyzed subsets. A condition for a model to be considered as performing well was to have an R^2 score ≥ 0.70 on the test set and to show no signs of overfitting, as evidenced by a variation in the R^2 score < 0.10 between the training and validation sets. These criteria allowed for an accurate and realistic analysis.

Moreover, the most important features were determined using the permutation importance method, followed by an aggregation of values, thus allowing the extraction of the 10 most relevant traits for each domain and country. Among the main conclusions is the fact that tree-based models (such as XGBoost, LightGBM, and Gradient Boosting) generally outperformed linear models, especially in Romania, France, and Hungary. However, in some cases, models like Ridge or ElasticNet showed competitive performance, highlighting the relevance of both approaches.

As observed, each subset had different models that performed better than others, but with a very small difference of approximately 0.03. However, looking at the overall picture, almost all six selected models had R^2 scores between 0.7 and 0.80, with linear regression standing out with two scores above 0.80 for Romania and 0.83 for Hungary. If we were to mention a model that performed well in most countries across all three domains, it would be ElasticNet. This model does not show overfitting, generalizes very well on unseen data, and also has R^2 scores above 0.70 in most subsets. Elastic Net performed best on the PISA 2022 subsets because it combines the advantages of two popular regularization methods: Lasso (L1) and Ridge (L2), shown in Appendix 2. Each subset may have small sample sizes, multicollinearity between variables, or noise. In such situations, Elastic Net is theoretically favored over other models.

Ridge tends to shrink coefficients but does not eliminate variables, which makes it useful when predictors are correlated. Lasso can completely eliminate unimportant variables, promoting sparse models, but becomes unstable when predictors are highly correlated. Elastic Net combines both: it retains Lasso’s ability to select variables while also benefiting from Ridge’s stability and regularization.

This flexibility is important for the PISA 2022 subsets, where variables can be redundant or strongly correlated. Elastic Net helps prevent overfitting and adapts to structural variations between countries, offering better generalization than more rigid or noise-sensitive models.

To sum it up, the theory behind Elastic Net makes it robust and adaptable. These qualities are essential when analyzing subsets like those in PISA 2022.

Also, the models performed worse on 3 of the 9 subsets for reading and science, these being Bulgaria, Estonia, and Serbia. This could be due to the fact that the PISA 2022 dataset was focused on mathematics, with more questions related to this domain. It is also important to mention that in the PISA 2022 cycle, the main tested domain was mathematics, which led to a wider coverage of items, higher scoring precision, and better predictive performance of the models in this area. In contrast, for Reading and Science, the smaller number of items and the lower variation of responses might negatively affect both the quality of the scores and the predictions generated by the models. This is reflected in the lower R^2 scores obtained by the models applied to these two domains.

Furthermore, a cosine similarity analysis between countries revealed similarities in the profiles of certain countries. For example, Bulgaria, Serbia, and Romania had very similar importance profiles across all three domains. Also, Estonia and Singapore, France with Hungary and Romania are other pairs of countries that present similar profiles. The country pairs considered were those with a similarity greater than 0.5 out of 1, the highest similarity being 0.62 between Romania and Bulgaria.

Moreover, at the trait level, variables such as the number of books at home (ST255Q01JA), the student's perception of family status (ST259Q01JA), parents' education level (ISCEDP), number of devices in the home (ST253Q01JA), time spent studying (ST059Q02JA), student involvement in physical activities (ST294Q05JA), exposure to contemporary literary resources (ST256Q03JA), subjective perception of mathematics difficulty (ST268Q04JA), and personal preference for mathematics (ST268Q01JA) consistently appeared as significant predictors.

Therefore, the frequency analysis of the appearance of questions such as "Mathematics is easy for me" or "Mathematics is one of my favorite subjects" in the predictive models used in this research highlights an essential aspect in understanding student performance: the role of perceptions and self-confidence. Indeed, these answers do not directly measure mathematical competence but reflect how students relate to this subject, which has a visible impact on the scores obtained.

Thus, students' confidence in their abilities, known in the specialized literature as self-efficacy, emerges as a determining factor in achieving good results in mathematics. Students who believe they can handle math tasks more easily are more likely to put in constant effort, try to solve more difficult problems, and maintain their motivation in the long term. Moreover, the fact that some students declare mathematics as one of their favorite subjects indicates a real interest in the field, which supports a more active and effective learning process.

On the other hand, even in situations where a student has a negative perception of their own abilities, they may perform below their actual potential. This suggests that educational interventions should not only focus on consolidating theoretical content but also on building a positive self-image among students. Promoting a favorable attitude towards mathematics and supporting confidence in success can play a decisive role in improving outcomes.

Therefore, these results support the idea that the socio-economic environment has a significant impact on academic performance, in line with the specialized literature. The paper also acknowledges several limitations: only nine countries were analyzed, which may reduce the generalizability of the conclusions; only six models were used for feature importance aggregation; and the data relied

exclusively on student questionnaires, without including information from schools or teachers, which could bring an additional perspective.
Beyond these aspects, the research opens the possibility for the development of personalized educational tools. The identified features can be used to anticipate the risk of school underperformance, offering early support to students. The trained models can be integrated into interactive applications that provide personalized feedback and predictions for students, parents, and teachers.
In conclusion, this paper successfully combined machine learning methods with educational data analysis, providing both theoretical insights and practical applicability. The results confirm the potential of predictive modeling in education and offer a flexible framework that can be extended to other countries, variables, or future PISA editions.

7.2 Future work

In future developments of this project, we can extend the analysis to include all countries participating in the PISA initiative. The current work focused on only 9 countries, but going forward, it would be beneficial to analyze all 81 participating countries to examine whether the selected models remain relevant for other subsets. Another important direction could be the exploration of new deep learning algorithms. In the current study, we used tree-based and regression models, but incorporating neural networks might enhance prediction accuracy, particularly for more complex relationships.
With the inclusion of neural networks, a longitudinal analysis could also be carried out, comparing data from PISA 2018 and PISA 2022 to assess the evolution of student performance over time and identify the most important predictive features for each country. This raises an essential question whether the same features remain relevant across multiple editions of the PISA assessment.
Additionally, as we plan to apply these models to over 80 countries, we could automate the feature selection process using methods such as Recursive Feature Elimination or mutual information to obtain more optimal subsets of variables. Another significant aspect for future work would be investigating cultural and curricular differences between countries. As observed in our analysis of the 9 countries, model performance varied considerably across contexts. These differences could be better understood by considering structural educational factors.
To make these insights accessible to a wider audience, including those outside the academic field, a user-friendly interface could be implemented. This would allow teachers or researchers to input data about a student and receive an estimated score prediction, along with the main features contributing to that prediction.

Bibliography

[0] Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. Journal of Development Economics, 104, 184–198. https://doi.org/10.1016/j.jdeveco.2012.10.001

[1] OECD, PISA 2022 Database, available at: https://www.oecd.org/en/data/datasets/pisa-2022-database.html (accessed on: February 6, 2025)

[2] Schmidt, A. M., de Moraes, C. P., & Migon, H. S. (2015). A Hierarchical Dynamic Beta Regression Model of School Performance in the Brazilian Mathematical Olympiads for Public Schools. https://doi.org/10.48550/arXiv.1507.00565

[3] Pavlik, P. I., & Eglington, L. G. (2021). Modeling the EdNet Dataset with Logistic Regression. https://doi.org/10.48550/arXiv.2105.08150

[4] Masci, C., Johnes, G., & Agasisti, T. (2018). Student and school performance across countries: A machine learning approach. European Journal of Operational Research, 269(3), 1072–1085. https://doi.org/10.1016/j.ejor.2018.02.031

[5] Pan, Z., & Cutumisu, M. (2024). Using machine learning to predict UK and Japanese secondary students' life satisfaction in PISA 2018. British Journal of Educational Psychology, 94(2), 474–498. https://doi.org/10.1111/bjep.12657

[6] Zhang, J., Mucs, D., Norinder, U., & Svensson, F. (2019). LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity-Application to the Tox21 and Mutagenicity Data Sets. Journal of Chemical Information and Modeling. https://doi.org/10.1021/acs.jcim.9b00633

[7] Seppänen, P., Rinne, R., Kauko, J., & Kosunen, S. (2019). The Use of PISA Results in Education Policy-Making in Finland. In Understanding PISA's Attractiveness. Bloomsbury Academic. https://doi.org/10.5040/9781350057319.ch-007

[8] Gabriel, F., Signolet, J., & Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. International Journal of Research & Method in Education, 41(3), 306–327. https://doi.org/10.1080/1743727X.2017.1301916

[9] Morales, E. F., & Escalante, H. J. (2022). A brief introduction to supervised, unsupervised, and reinforcement learning. In Biosignal Processing and Classification Using Computational Learning and Intelligence (pp. 111–129). Elsevier. https://doi.org/10.1016/B978-0-12-820125-1.00017-8

[10] Alpaydın, E. (2021). Machine Learning. The MIT Press. https://doi.org/10.7551/mitpress/13811.001.0001

[11] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, Foundations of Machine Learning. MIT Press, Second Edition, 2018.

[12] Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. Water Resources Management, 31(9), 2761–2775. https://doi.org/10.1007/s11269-017-1660-3

[13] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

[14] Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest (Vol. 2, Issue 3). http://www.stat.berkeley.edu/

[15] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (n.d.). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. https://github.com/Microsoft/LightGBM

[16] Anghel, A., Papandreou, N., Parnell, T., de Palma, A., & Pozidis, H. (2018). Benchmarking and Optimization of Gradient Boosting Decision Tree Algorithms. http://arxiv.org/abs/1809.04559

[17] Florek, P., & Zagdański, A. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification. http://arxiv.org/abs/2305.17094

[18] Salvador, E. L. (2024). Use of Boosting Algorithms in Household-Level Poverty Measurement: A Machine Learning Approach to Predict and Classify Household Wealth Quintiles in the Philippines. http://arxiv.org/abs/2407.13061
[19] Chen, T., & Guestrin, C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
[20] Shahani, N. M., Zheng, X., Liu, C., Hassan, F. U., & Li, P. (2021). Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures. Frontiers in Earth Science, 9. https://doi.org/10.3389/feart.2021.761990
[21] Wen, H.-T., Wu, H.-Y., & Liao, K.-C. (2022). Using XGBoost Regression to Analyze the Importance of Input Features Applied to an Artificial Intelligence Model for the Biomass Gasification System. Inventions, 7(4), 126. https://doi.org/10.3390/inventions7040126
[22] Hafid, A., Ebrahim, M., Alfatemi, A., Rahouti, M., & Oliveira, D. (2024). Cryptocurrency Price Forecasting Using XGBoost Regressor and Technical Indicators. http://arxiv.org/abs/2407.11786
[23] Zemel, R. S., & Pitassi, T. (n.d.). A Gradient-Based Boosting Algorithm for Regression Problems.
[24] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021
[25] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785
[26] Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 15(1), 73. https://doi.org/10.1186/s13321-023-00743-7
[27] Seto, H., Oyama, A., Kitora, S., Toki, H., Yamamoto, R., Kotoku, J., Haga, A., Shinzawa, M., Yamakawa, M., Fukui, S., & Moriyama, T. (2022). Gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. Scientific Reports, 12(1), 15889. https://doi.org/10.1038/s41598-022-20149-z
[28] Wang, W., & Zhou, Z.-H. (2008). On multi-view active learning and the combination with semi-supervised learning. Proceedings of the 25th International Conference on Machine Learning - ICML '08, 1152–1159. https://doi.org/10.1145/1390156.1390301
[29] Yavari, H. (2024). Solution gas-oil ratio estimation using histogram gradient boosting regression, machine learning, and mathematical models: a comparative analysis. Energy Sources, Part A: Recovery, Utilization and Environmental Effects, 46(1), 379–396. https://doi.org/10.1080/15567036.2023.2284236
[30] Jaafar, A. G., Ismail, S. A., Habir, A., Ariffin, K. A. Z., & Yusop, O. M. (2024). A Raise of Security Concern in IoT Devices: Measuring IoT Security Through Penetration Testing Framework. International Journal of Advanced Computer Science and Applications, 15(5). https://doi.org/10.14569/IJACSA.2024.0150568
[31] Juckem, P. F., Corson-Dosch, N. T., Schachter, L. A., Green, C. T., Ferin, K. M., Booth, E. G., Kucharik, C. J., Austin, B. P., & Kauffman, L. J. (2024). Design and calibration of a nitrate decision support tool for groundwater wells in Wisconsin, USA. Environmental Modelling & Software, 176, 105999. https://doi.org/10.1016/j.envsoft.2024.105999
[32] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression Diagnostics. Wiley. https://doi.org/10.1002/0471725153
[33] Schmidt, A. M., de Moraes, C. P., & Migon, H. S. (2015). A Hierarchical Dynamic Beta Regression Model of School Performance in the Brazilian Mathematical Olympiads for Public Schools. https://doi.org/10.48550/arXiv.1507.00565
[34] Yuan, Z., & Yang, Y. (2005). Combining linear regression models: When and how? Journal of the American Statistical Association, 100(472), 1202–1214. https://doi.org/10.1198/016214505000000088

[35] Li, J., Yang, T., & Zheng, R. (2024). Theoretical simulation of the structure–activity relationship of polyimide dielectric constant and analysis of its linear regression model. Applied Physics A: Materials Science and Processing, 130(1). https://doi.org/10.1007/s00339-023-07238-0
[36] Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied Linear Regression Models (4th ed.).
[37] Harrell,, Frank E. (2015). Regression Modeling Strategies. Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7
[38] Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634
[39] van Wieringen, W. N. (2015). Lecture notes on ridge regression. https://doi.org/10.48550/arXiv.1509.09169
[40] Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12(1), 69–82. https://doi.org/10.1080/00401706.1970.10488635
[41] Willoughby, R. A. (1979). Solutions of Ill-Posed Problems (A. N. Tikhonov and V. Y. Arsenin). SIAM Review, 21(2), 266–267. https://doi.org/10.1137/1021044
[42] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-84858-7
[43] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[44] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer New York. https://doi.org/10.1007/978-0-387-84858-7
[45] Bühlmann, P., & van de Geer, S. (2011). Statistics for High-Dimensional Data. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20192-9
[46] Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. The Annals of Statistics, 34(3). https://doi.org/10.1214/009053606000000281
[47] Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
[48] Park, M. Y., & Hastie, T. (2007). L 1-Regularization Path Algorithm for Generalized Linear Models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 69(4), 659–677. https://doi.org/10.1111/j.1467-9868.2007.00607.x
[49] Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis, 5(2). https://doi.org/10.1214/10-BA607
[50] Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
[51] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1). https://doi.org/10.18637/jss.v033.i01
[52] de Mol, C., de Vito, E., & Rosasco, L. (2008). Elastic-Net Regularization in Learning Theory. https://doi.org/10.48550/arXiv.0807.3423
[53] Park, S., Kim, W., & Lee, K. M. (2012). Abnormal Object Detection by Canonical Scene-Based Contextual Model (pp. 651–664). https://doi.org/10.1007/978-3-642-33712-3_47
[54] Molineaux, M., & Aha, D. (2014). Learning Unknown Event Models. Proceedings of the AAAI Conference on Artificial Intelligence, 28(1). https://doi.org/10.1609/aaai.v28i1.8751
[55] PISA 2022 Assessment and Analytical Framework. Retrieved from https://www.oecd.org/en/publications/pisa-2022-assessment-and-analytical-framework_dfe0bf9c-en.html (accessed on: February 27, 2025).

[56] Schweri, J., Hartog, J., & Wolter, S. C. (2011). Do students expect compensation for wage risk? Economics of Education Review, 30(2), 215–227. https://doi.org/10.1016/j.econedurev.2010.12.001
[57] PISA 2022 Database, available at: https://www.oecd.org/en/data/datasets/pisa-2022-database.html (accessed on: February 27, 2025).
[58] Creative Thinking and Financial Literacy in PISA 2022, available at: https://www.cmec.ca/712/PISA_2022.html (accessed on: February 27, 2025).
[59] Technical Standards for PISA Sampling, available at: https://nces.ed.gov/surveys/pisa/pisa2022/ (accessed on: February 27, 2025).
[60] Education Recovery Post-Pandemic: Lessons from PISA, available at: https://en.unesco.org (accessed on: February 27, 2025).
[61] Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2013). Handbook of International Large-Scale Assessment. Chapman and Hall/CRC. https://doi.org/10.1201/b16061
[62] Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. Information Sciences, 307 https://doi.org/10.1016/j.ins.2015.02.024
[63] Zhu, L., You, H., Hong, M., & Fang, Z. (2025). Predictive insights into U.S. students' mathematics performance on PISA 2022 using ensemble tree-based machine learning models. International Journal of Educational Research, 130, 102537. https://doi.org/10.1016/j.ijer.2025.102537
[64] Huang, Y., Zhou, Y., Chen, J., & Wu, D. (2024). Applying Machine Learning and SHAP Method to Identify Key Influences on Middle-School Students' Mathematics Literacy Performance. Journal of Intelligence, 12(10), 93. https://doi.org/10.3390/jintelligence12100093
[65] Öz, E., Bulut, O., Cellat, Z. F., & Yürekli, H. (2025). Stacking: An ensemble learning approach to predict student performance in PISA 2022. Education and Information Technologies, 30(6), 7753–7779. https://doi.org/10.1007/s10639-024-13110-2
[66] PISA 2022 Results (Volume I). (2023). OECD. https://doi.org/10.1787/53f23881-en
[67] How to prepare and analyse the PISA database https://www.oecd.org/en/about/programmes/pisa/how-to-prepare-and-analyse-the-pisa-database.html (accessed on: February 6, 2025)
[68] Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating Population Characteristics From Sparse Matrix Samples of Item Responses. Journal of Educational Measurement, 29(2), 133–161. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Appendix

This appendix does not present the complete cade base behind the project but highlights essential segments that are central to the analysis and implementation.

Appendix 1

```
from IPython.core.magic import register_cell_magic

@register_cell_magic
def skip(line, cell):
    return

import numpy as np
import pandas as pd
import pyreadstat
import joblib

import matplotlib.pyplot as plt
import seaborn as sb
from sklearn.pipeline import Pipeline
from sklearn.utils import shuffle

import seaborn as sns
import numpy as np


df = pd.read_parquet('CY08MSP_STU_QQQ.SAV_old.parquet', engine='pyarrow')
# Add mean PV values in each category: MATH, READ, SCI, and the 8 sub-
categories of MATH

#Plausible Value in Mathematics
df['PVMATH'] = df[['PV1MATH', 'PV2MATH', 'PV3MATH', 'PV4MATH', 'PV5MATH',
                   'PV6MATH', 'PV7MATH', 'PV8MATH', 'PV9MATH', 'PV10MATH',
                  ]].mean(axis=1)

#Plausible Value in Reading
df['PVREAD'] = df[['PV1READ', 'PV2READ', 'PV3READ', 'PV4READ', 'PV5READ',
                   'PV6READ', 'PV7READ', 'PV8READ', 'PV9READ', 'PV10READ',
                  ]].mean(axis=1)

#Plausible Value in Science
df['PVSCIE'] = df[['PV1SCIE', 'PV2SCIE', 'PV3SCIE', 'PV4SCIE', 'PV5SCIE',
                   'PV6SCIE', 'PV7SCIE', 'PV8SCIE', 'PV9SCIE', 'PV10SCIE',
                  ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Change and
Relationships
df['PVMCCR'] = df[['PV1MCCR', 'PV2MCCR', 'PV3MCCR', 'PV4MCCR', 'PV5MCCR',
```

```
                    'PV6MCCR', 'PV7MCCR', 'PV8MCCR', 'PV9MCCR', 'PV10MCCR',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Quantity
df['PVMCQN'] = df[['PV1MCQN', 'PV2MCQN', 'PV3MCQN', 'PV4MCQN', 'PV5MCQN',
                    'PV6MCQN', 'PV7MCQN', 'PV8MCQN', 'PV9MCQN', 'PV10MCQN',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Space and Shape
df['PVMCSS'] = df[['PV1MCSS', 'PV2MCSS', 'PV3MCSS', 'PV4MCSS', 'PV5MCSS',
                    'PV6MCSS', 'PV7MCSS', 'PV8MCSS', 'PV9MCSS', 'PV10MCSS',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Uncertainty and Data
df['PVMCUD'] = df[['PV1MCUD', 'PV2MCUD', 'PV3MCUD', 'PV4MCUD', 'PV5MCUD',
                    'PV6MCUD', 'PV7MCUD', 'PV8MCUD', 'PV9MCUD', 'PV10MCUD',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Employing Mathematical
Concepts, Facts, and Procedures
df['PVMPEM'] = df[['PV1MPEM', 'PV2MPEM', 'PV3MPEM', 'PV4MPEM', 'PV5MPEM',
                    'PV6MPEM', 'PV7MPEM', 'PV8MPEM', 'PV9MPEM', 'PV10MPEM',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Formulating Situations
Mathematically
df['PVMPFS'] = df[['PV1MPFS', 'PV2MPFS', 'PV3MPFS', 'PV4MPFS', 'PV5MPFS',
                    'PV6MPFS', 'PV7MPFS', 'PV8MPFS', 'PV9MPFS', 'PV10MPFS',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Interpreting,
Applying, and Evaluating Mathematical Outcomes
df['PVMPIN'] = df[['PV1MPIN', 'PV2MPIN', 'PV3MPIN', 'PV4MPIN', 'PV5MPIN',
                    'PV6MPIN', 'PV7MPIN', 'PV8MPIN', 'PV9MPIN', 'PV10MPIN',
                   ]].mean(axis=1)

#Plausible Value in Content Subscale of Mathematics - Reasoning
df['PVMPRE'] = df[['PV1MPRE', 'PV2MPRE', 'PV3MPRE', 'PV4MPRE', 'PV5MPRE',
                    'PV6MPRE', 'PV7MPRE', 'PV8MPRE', 'PV9MPRE', 'PV10MPRE',
                   ]].mean(axis=1)

categ_names = [
'PVMATH',               #Plausible Value 1 in Mathematics
'PVREAD',               #Plausible Value 1 in Reading
'PVSCIE',               #Plausible Value 1 in Science
'PVMCCR',               #Plausible Value 1 in Content Subscale of
Mathematics - Change and Relationships
'PVMCQN',               #Plausible Value 1 in Content Subscale of
Mathematics - Quantity
'PVMCSS',               #Plausible Value 1 in Content Subscale of
Mathematics - Space and Shape
```

```
'PVMCUD',               #Plausible Value 1 in Content Subscale of
Mathematics - Uncertainty and Data
'PVMPEM',               #Plausible Value 1 in Process Subscale of
Mathematics - Employing Mathematical Concepts, Facts, and Procedures
'PVMPFS',               #Plausible Value 1 in Process Subscale of
Mathematics - Formulating Situations Mathematically
'PVMPIN',               #Plausible Value 1 in Process Subscale of
Mathematics - Interpreting, Applying, and Evaluating Mathematical Outcomes
'PVMPRE',               #Plausible Value 1 in Process Subscale of
Mathematics - Reasoning
]



#Calculates the number of null values for each section
#as well as percent of students who answered all questions under each
section
i=0
missing_data=[]
f=0
for _ in df:
    i+=1
    d=df[_].isnull().sum()
    p=d/df.shape[0]

    if p > 0.50:
        print(i)
        f=f+1
        missing_data.append((_,p))
        print(_, "missing:", d)
        print('% missing: {:.0%}'.format(d/df.shape[0]))
        print('\n')


print(f)
# Reset the seed of the random number generator, for reproducibility
purposes

import os

def reset_seed(SEED = 0):
    """Reset the seed for every random library in use (System, numpy)"""

    os.environ['PYTHONHASHSEED']=str(SEED)
    np.random.seed(SEED)


reset_seed(2024)


# Define a function to split the data into train/validation/test
```

```
# This will be called several times in different setups, it simplifies
reading the code

def data_split(X):
    X_train_valid, X_test = train_test_split(
        X,
        test_size=0.2,
        random_state=150,
        shuffle=True
    )

    X_train, X_valid = train_test_split(
        X_train_valid,
        test_size=0.25,
        random_state=150,
        shuffle=True
    )

    # convert to pandas dataframe
    X_train = pd.DataFrame(X_train, columns=X.columns)
    X_valid = pd.DataFrame(X_valid, columns=X.columns)
    X_test = pd.DataFrame(X_test, columns=X.columns)

    return X_train, X_valid, X_test
from sklearn.inspection import permutation_importance

def my_permutation_importance(models, df_valid, target, pipeline):

    y_valid = df_valid['PV'+target]
    X_valid = df_valid[ y_valid.notna() ].drop(scores + categ_names, axis=1)
    y_valid = y_valid[ y_valid.notna() ]


    X_valid_scaled = pipeline.transform(X_valid)


    importances_mean = np.zeros(X_valid_scaled.shape[1])
    importances_std = np.zeros(X_valid_scaled.shape[1])

    for i in range(1, 11):
        r = permutation_importance(models[i-1], X_valid_scaled, y_valid,
n_repeats=5, random_state=2024, n_jobs=-1)
        importances_mean += r.importances_mean
        importances_std += r.importances_std
    importances_mean /= 10
    importances_std /= 10
    values=[]
    print("Most important permutation features on ", target, ":")
    for i in importances_mean.argsort()[::-1]:
        if importances_mean[i] - 2 * importances_std[i] > 0 and
importances_mean[i] > 0.001:
```

```
            name = X_valid.columns[i]
            importance=importances_mean[i]
            std=importances_std[i]
            print(f"{name:<30} {importance:.4f} ± {std:.4f}")
            values.append((importance, name))
    return (importances_mean, values)
```

Appendix 2

```
all_data={}

from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import ElasticNet

def train_ElasticNet_reg(df_train, df_valid, df_test, target=None):

    models = [None] * 10
    pipeline = Pipeline([
        ('imputer', SimpleImputer(strategy='most_frequent')),
        ('scaler', StandardScaler())
         ])

    for i in range(1, 11):
        # Train on PVi, cu i = 1, 2, ..., 10
        current_target = 'PV' + str(i) + target
        print("\t Training on ", current_target)
        y_train = df_train[current_target]

        # Drop features not wanted and select non-NA values
        X_train = df_train[y_train.notna()].drop(scores + categ_names,
axis=1)
        y_train = y_train[y_train.notna()]

        # Impute and scale data
        X_train_scaled = pipeline.fit_transform(X_train)

        # Train model
        reg = ElasticNet( alpha=1, l1_ratio=0.5

                    )
        reg.fit(X_train_scaled, y_train)
        models[i-1] = reg

    # Preprocess validation data
    y_train = df_train['PV' + target]
```

```
    X_train = df_train[y_train.notna()].drop(scores + categ_names, axis=1)
    y_train = y_train[y_train.notna()]
    y_valid = df_valid['PV' + target]
    X_valid = df_valid[y_valid.notna()].drop(scores + categ_names, axis=1)
    y_valid = y_valid[y_valid.notna()]
    y_test = df_test['PV' + target]
    X_test = df_test[y_test.notna()].drop(scores + categ_names, axis=1)
    y_test = y_test[y_test.notna()]

    # Apply imputation and scaling to the validation data
    X_train_scaled = pipeline.transform(X_train)
    X_valid_scaled = pipeline.transform(X_valid)
    X_test_scaled = pipeline.transform(X_test)

    # Average of the predictions from each model
    y_train_pred = [0] * X_train.shape[0]
    y_valid_pred = [0] * X_valid.shape[0]
    y_test_pred = [0] * X_test.shape[0]
    for i in range(1, 11):
        y_train_pred += models[i-1].predict(X_train_scaled)
        y_valid_pred += models[i-1].predict(X_valid_scaled)
        y_test_pred += models[i-1].predict(X_test_scaled)
    y_train_pred /= 10
    y_valid_pred /= 10
    y_test_pred /= 10
    #calculate metrics
    rmse_train = mean_squared_error(y_train, y_train_pred, squared=False)
    rmse_valid = mean_squared_error(y_valid, y_valid_pred, squared=False)
    rmse_test = mean_squared_error(y_test, y_test_pred, squared=False)


    # Diplays the metrics and the feature importance
    print(f'Train R2: {R2(y_train, y_train_pred):.2f}')
    print(f'Validation R2: {R2(y_valid, y_valid_pred):.2f}')
    print(f'Test R2: {R2(y_test, y_test_pred):.2f}')
    print(f'Train RMSE: {rmse_train:.2f}')
    print(f'Validation RMSE: {rmse_valid:.2f}')
    print(f'Test RMSE: {rmse_test:.2f}')

    print('Calculating permutation importance: ')
    importance, name = my_permutation_importance(models, df_valid,
target,pipeline)

    all_data[f'{type(reg).__name__}_{target}']=name

    return models, y_test, y_test_pred



X_train, X_valid, X_test = data_split(X)
```

```
all=[ 'MATH', 'READ', 'SCIE' ]


for index in all:
     print(f"Training for the category: {index}")
     train_ElasticNet_reg(X_train, X_valid, X_test, target=index)
```