Data Science Final Project: Predicting Mathematics Self-Efficacy using the PISA 2022 Dataset

Lucas Salamanca

November 2025

1 Introduction

The goal of this project is to predict students' mathematics self-efficacy (MATHEFF) using a wide set of individual, family, and school-level characteristics from the PISA 2022 dataset. Many people grow up disliking math throughout their school years, often because they struggle with the subject or feel like they are "bad" at it. Personally, I believe math can be genuinely enjoyable, but it usually requires good and inspiring teachers, which is something that, unfortunately, is not always common. In the PISA framework, mathematics self-efficacy refers to how confident students feel in their ability to solve mathematical problems. This is an important non-cognitive skill that shapes engagement, motivation, persistence, and ultimately academic performance. Being able to predict self-efficacy from observable characteristics is useful because it can help schools and policymakers identify which factors are most strongly associated with students' beliefs about their mathematical abilities, and which groups may need targeted support.

The problem is formulated as a regression task, since MATHEFF is a continuous index. To approach this, I selected the entire set of conceptual predictors from the PISA dataset, from both the student and school questionnaires. Conceptual predictors are both ready-made continuous indexes and categories built from the multiple Likert-scale questions that the students actually answer during the test. This set already includes commonly studied predictors in educational research, including psychological constructs (e.g., anxiety, perseverance, creativity), socioeconomic indicators (e.g., parental education, household possessions), school climate measures, and variables describing classroom practices and learning environments.

The modelling has two main objectives. First, we want to build predictive models that perform well on unseen data and compare them across traditional statistical methods and modern machine-learning approaches. For this reason, we estimate a baseline Ordinary Least Squares (OLS) model, several regularised linear models (Ridge, LASSO, Elastic Net), and two tree-based

machine learning algorithms (Random Forest and XGBoost). Second, we want to identify which variables contribute most to predicting MATHEFF. Regularised regression allows us to assess which predictors retain explanatory power after penalisation, which is relevant in our high-dimensionality setting. On the other hand, Random Forest and XGBoost provide measures of variable importance capturing nonlinear effects and interactions. This is particularly relevant for efficient policy targeting, as mentioned earlier.

2 Data

The data used in this project comes from the PISA 2022 international assessment, administered by the OECD. The dataset contains information on 15-year-old students from 74 countries, collected through standardized cognitive tests and detailed background questionnaires. Each row in the dataset represents a single student. The analysis combines variables from both the student questionnaire and the school questionnaire, merged using a common school identifier.

This initial merged dataset contained 613,744 observations, each of them representing a student, and 1,708 variables, which correspond to conceptual variables, original Likert-scale questions, IDs, sample weights, questionnaire-specific flags and raw occupation and educational codes. In total there are 114 conceptual variables that are built to capture the information from the original Likert-scale questions in a standardized way and to avoid redundancy. Thus, these conceptual variables will be the ones used for my analysis.

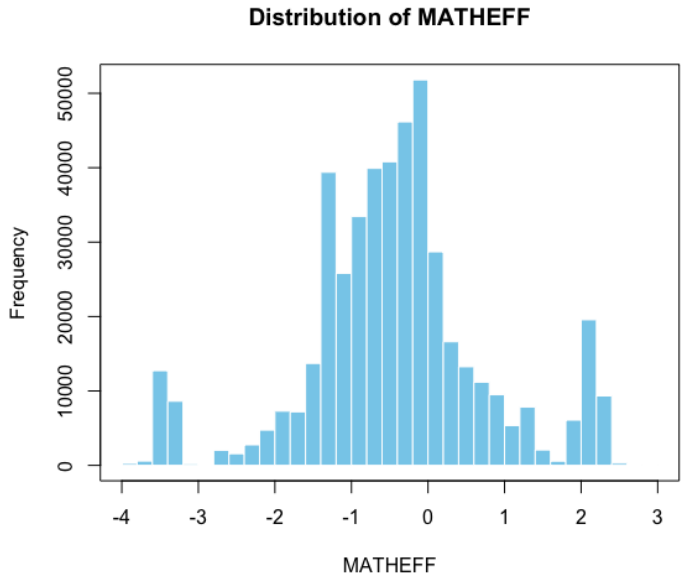

Figure 1: Histogram of MATHEFF

Regarding my target variable, MATHEFF, it is a continuous index of mathematics self-efficacy, scaled by the OECD. In Figure 1 we can observe how MATHEFF resembles a bell-shaped distribution with little asymmetry, which means that the variable is suitable for regression modelling without transformation. This distribution also suggests that extreme values are rare, reducing concerns about heavy tails.

I also examined correlations between the tentative regressors and MATHEFF. The heatmap presented in Figure 2 highlights the twenty variables most strongly correlated with MATHEFF. The correlations are moderate in size, ranging from approximately 0.16 to 0.42. The strongest correlated variables are FAMCON (subjective familiarity with mathematics concepts), MATHPERS (persistence in mathematics), and ANXMAT (mathematics anxiety). Also, the direction of the correlations matches the theoretical expectations: variables related to perseverance, curiosity, cognitive strategies, and supportive family or school environments show positive associations with mathematics self-efficacy, whereas anxiety in mathematics (ANXMAT) shows a clear negative association.

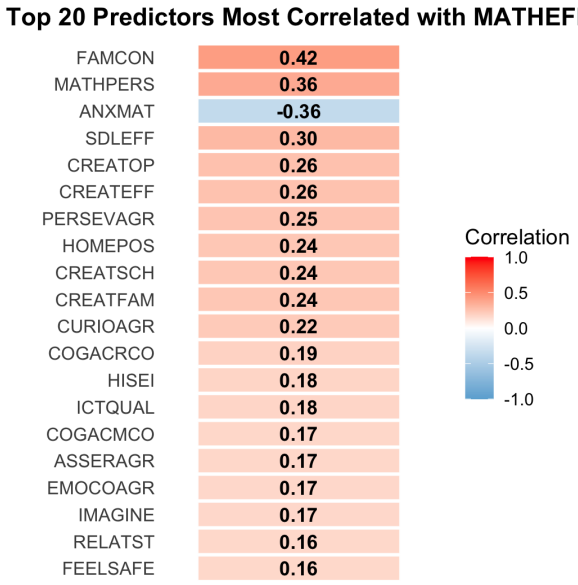

Figure 2: Correlation heatmap (top 20 variables correlated with MATHEFF)

Although these correlations provide a useful initial ranking of influential predictors, they are limited to linear and pairwise relationships. In that sense, to understand whether the relationships with MATHEFF are adequately captured by linear terms or whether more flexible modeling is required, we examine scatterplots with loess smoothers for the highest-ranked predictors.

The plots presented in Figure 2 show that ANXMAT has a clearly nonlinear negative relationship, MATHPERS shows a positive association with diminishing returns, and FAMCON seems mostly

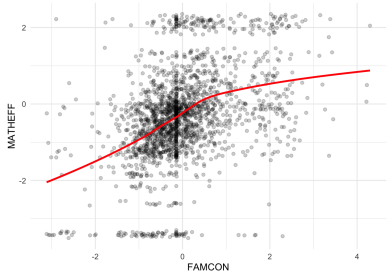

Figure 3: * FAMCON vs. MATHEFF

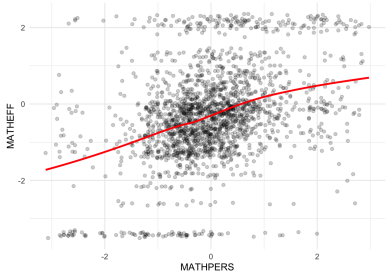

Figure 4: * MATHPERS vs. MATHEFF

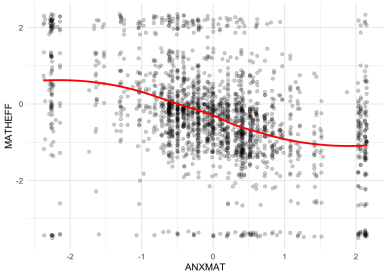

Figure 5: * ANXMAT vs. MATHEFF

linear but still shows curvature around the center and a plateau at higher values. Together, these patterns highlight the need for models capable of capturing both linear and nonlinear effects. Thus, flexible tree-based models (Random Forest and XGBoost) are well suited for this prediction task.

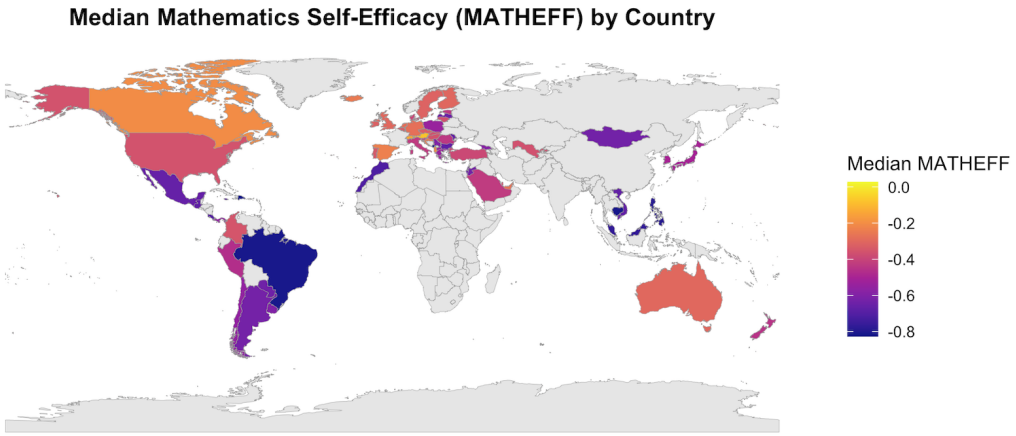

Figure 6: Median MATHEFF by country

Finally, Figure 6 shows that there are clear cross-country differences in median MATHEFF. Countries such as Australia, Canada, and parts of Western Europe show higher levels of mathematics self-efficacy, while many Latin American and Southeast Asian countries show lower values. These differences suggest that national economic and educational contexts contribute to variation in MATHEFF scores, reinforcing the value of using flexible tree-based models that can naturally account for complex, heterogeneous patterns across countries.

3 Data Cleaning and Processing

The starting point for the analysis were the official SPSS files from PISA 2022: the student questionnaire and the school questionnaire. The two files were then merged using the international school identifier CNTSCHID, performing a left join that keeps all students and attaches the characteristics of their school.

Because PISA includes thousands of item-level variables (individual questionnaire questions), I first restricted the analysis to conceptual indices instead of raw items. Using the merged variable dictionary, I kept variables whose names consist only of uppercase letters and have length ≥ 5, which corresponds to indices such as ANXMAT, MATHPERS, HOMEPOS, etc., while automatically excluding item codes like ST341Q02JA. I also manually ensured that the outcome variable MATHEFF and a small set of key controls were included. This step allows for an initial reduction in dimensionality and focuses the analysis on variables that already aggregate several items into interpretable scales.

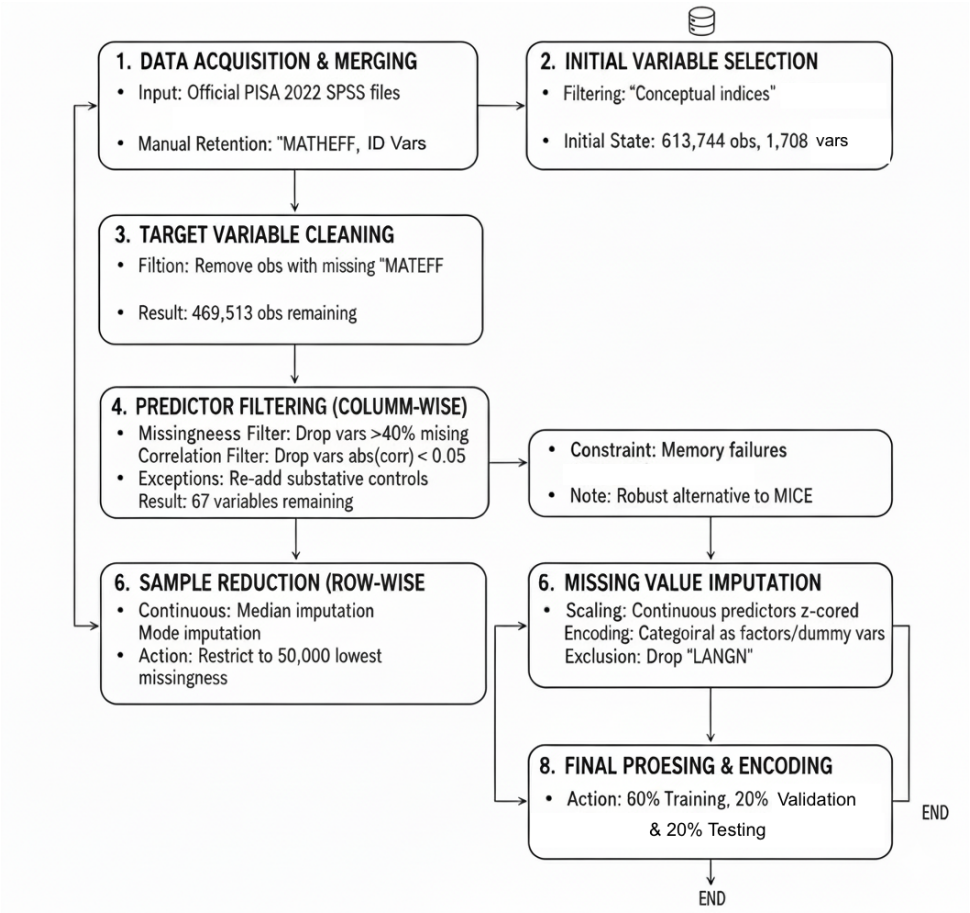

Figure 7: Data cleaning and processing pipeline

Next, I removed all observations with missing MATHEFF. Since MATHEFF is the target of prediction, these rows cannot be used for supervised learning and would only complicate the modelling. However, after removing these observations I was still left with 469,513 observations, which is more than sufficient for this analysis.

I then examined missing data for the rest of predictors. Variables with more than 40% missingness were dropped. Most of these variables represented country- and region-specific attributes, which justifies why they were also dropped. Another benefit of this decision is that it avoids spending imputation effort on variables with very little observed information and reduces the risk of imputations being driven mostly by modeling assumptions rather than data. The trade-off is that some potentially interesting but sparsely observed constructs are excluded. However, given the large number of remaining indices and the focus on prediction rather than exhaustive description, the process is methodologically fair.

The resulting dataframe had 67 variables. Before imputing missing values for these observations, I created a reduced sample of students with fewer missing entries. At first I tried to run the full set of predictive models on the complete dataset of 469,513 students, but all the regressions and tree-based models failed to run due to memory limits and extremely long computation times. For this reason, I restricted the modelling dataset to the 50,000 students with the lowest missingness. This keeps the sample still very large while focusing on cases with relatively complete information, which makes imputation more reliable and helps with the computational burden. A possible drawback is that this may slightly over-represent students from better-documented contexts (e.g., countries or schools with fewer skipped items), but given the overall size and diversity of the subsample this bias is limited.

To decide how to impute, I built a small dictionary listing for each variable its label and the set of distinct observed values. Using this, I classified variables into continuous (WLE scales, indices, counts) and categorical (binary indicators, ordered education levels, school type, immigrant status, etc.). For the imputation of continuous variables I imputed them with their median. This choice is robust to outliers and preserves the central tendency without being influenced by extreme values. For categorical variables I used mode imputation, which ensures that imputed values correspond to valid categories and roughly preserve the original distribution.

In fact, I originally attempted multivariate imputation via MICE, but the combination of large n and many predictors led to repeated "vector memory exhausted" errors. Given these computational constraints, median/mode imputation is a pragmatic alternative that avoids dropping a large proportion of students (only about 1% of rows are fully complete) and is easy to explain and reproduce.

4 Feature Engineering

I then prepared the final modelling dataset by removing identifier variables (student and school IDs, booklets) from the predictor set, since they have no predictive meaning. After that, in order to reduce the high dimensionality levels, variables with an absolute correlation below 0.05 were dropped from the modelling set. I also added back a set of substantively important controls (e.g., immigration background, parental education, grade repetition) regardless of their correlation. It is also relevant to note that variables that might matter only through interactions or nonlinear effects could be excluded, but in practice almost all psychologically and socioeconomically important indices survived the threshold.

Continuous predictors were standardized (z-scores) while leaving MATHEFF in its original scale. Standardisation is important for regularised regressions, which penalise coefficient magnitude because if the variables are on different scales, the penalty would unfairly shrink large-scale variables more. Scaling also makes coefficient sizes comparable across predictors.

Categorical variables remained as factors. For methods that require numeric input (glmnet and XGBoost), I used `sparse.model.matrix` to automatically create dummy variables for each factor level. I also excluded one very high-cardinality categorical variable (LANGN, language spoken at home) from the penalised and tree-based models to avoid an explosion in the number of dummy variables and associated memory use.

5 Methodology

For this project I estimated five models: a benchmark OLS (only using the top 20 correlated variables), three regularised models (Ridge, LASSO, and Elastic Net), and two tree-based models (Random Forest and XGBoost). Previous studies that use PISA data have shown that relationships between psychological constructs (e.g., anxiety, perseverance, cognitive strategies) and outcomes such as achievement or self-efficacy are often nonlinear, making tree-based models more accurate compared to linear models (Tan & Cutumisu, 2022; Tan, Jin, & Cutumisu, 2025). Additionally, XGBoost and Random Forest are also able to capture heterogeneous effects across schools, countries, and socioeconomic groups accurately.

However, using regularised models is also pertinent in this setting. Ridge addresses the strong multicollinearity among PISA indices, LASSO performs efficient variable selection in high-dimensional settings, and Elastic Net mixes both advantages, which is valuable when correlated predictors should be jointly retained. Finally, OLS is included as a baseline to quantify the predictive gain from the other models.

The dataset was randomly split into three mutually exclusive subsets: 60% for training, 20% for validation, and 20% for testing. The purpose of this split is to impose a strict methodological separation between (1) estimating model parameters, (2) choosing hyperparameters, and (3) evaluating final performance. This division avoids "peeking" at the test set during tuning and yields an unbiased estimate of real-world predictive performance.

Regarding the tuning of hyperparameters, for Ridge, LASSO, and Elastic Net, I used a validation-based tuning scheme, in line with what we have seen in the course. A logarithmic grid of 100 lambda values (10^{-4} to 10^{2}) was evaluated. Each model was fitted on the training set and then used to predict MATHEFF in the validation set, selecting the lambda that minimised the validation RMSE. This procedure ensures that hyperparameter choices depend only on validation performance.

For the Random Forest hyperparameters I tuned analogously by testing combinations of `mtry` ($\sqrt{p}$, $p/4$, $p/2$) and number of trees (100, 200), again selecting the configuration yielding the lowest validation error. However, for the RF it is relevant to note that due to computational constraints (yet again) I had to build a smaller subsample of 10,000 observations and limit the number of trees to 200, as otherwise the computer (MacBook Air 2019, unfortunately the model before the M1 chip) crashed. For XGBoost I used a small but effective hyperparameter grid (`max_depth` $\in \{4, 6\}$, $\eta \in \{0.05, 0.10\}$, `subsample` and `colsample_bytree` fixed at 0.8), with early stopping determined only by validation RMSE. This search structure mirrors the strategy implemented in published PISA ML analyses, which consistently emphasise that shallow boosted trees and moderate learning rates perform best with large-scale educational data.

6 Results

Metric	OLS (top 20)	Ridge	LASSO	Elastic Net	Random Forest	XGBoost
R^2	0.345084	0.398347	0.398848	0.398815	0.371522	0.419590
RMSE	0.955745	0.916056	0.915675	0.915700	0.936255	0.899739

Table 1: Test set performance metrics by model

The test-set results in Table 1 show relevant differences in predictive performance across the models. The best performing model is XGBoost, achieving the lowest RMSE and highest R^2, followed closely by the regularised regressions, while OLS and Random Forest lag behind. The regularised linear models perform well because many of the conceptual PISA indices have strong linear relationships with MATHEFF. However, the additional gains from XGBoost suggest that nonlinear interactions, particularly between motivational and affective variables, play an important role and can only be captured by flexible tree-based approaches.

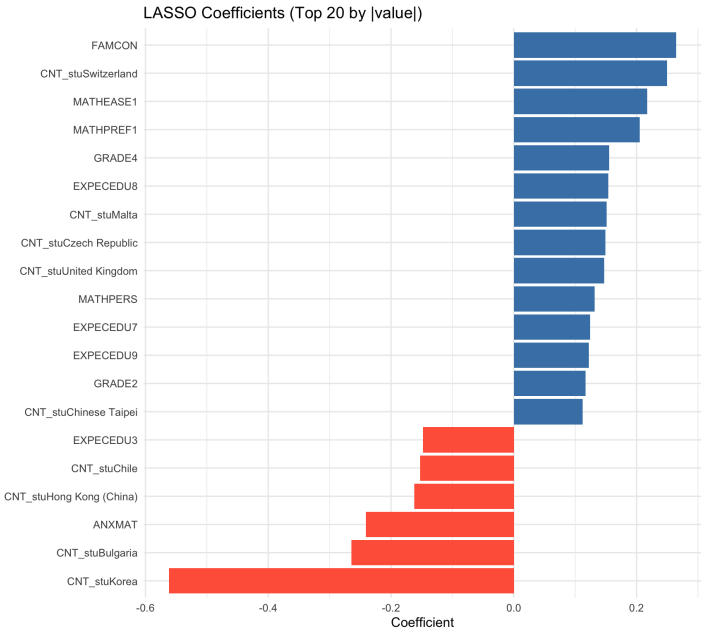

Figure 8: LASSO: Coefficient magnitudes after regularisation

Figure 9: Random Forest: Variable importance

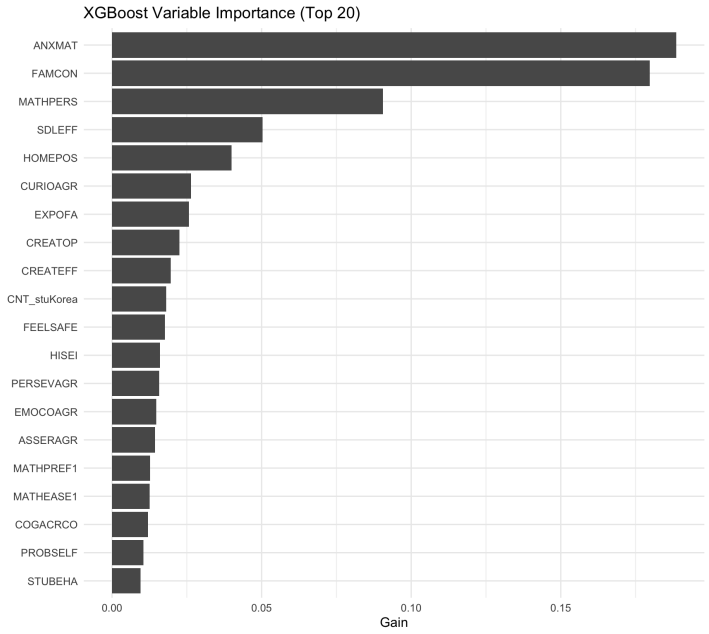

Figure 10: XGBoost: Variable importance

We can also observe that across all models FAMCON (Familiarity with Mathematics concepts), MATHPERS (Persistance in Mathematics), SDLEFF (Self-directed learning efficacy), PERSEVAGR (Perseverance), and the strongly negative ANXMAT (Mathematics anxiety) appear as the top predictors in LASSO, Random Forest, and XGBoost (Figures 8,9 and 10). The signs and magnitudes of the coefficients show that variables that reflect confidence, familiarity, and self-regulation reliably increase mathematics self-efficacy, whereas anxiety has an important negative effect. These findings align with educational psychology literature. It is also relevant to note the presence of creativity-related constructs in the tree-based models' importance rankings (e.g., CREATOP, CREATSCH), which suggests that school environments that promote creative thinking may interact with individual motivation, reinforcing self-efficacy beyond what linear models capture. Additionally, several country identifiers appear among the top predictors, which is consistent with the clear cross-country differences in MATHEFF highlighted earlier in the data section. This suggests that national context still plays a meaningful role in shaping students' self-efficacy, even after accounting for individual-level characteristics. Finally, an advantage of LASSO compared to the tree-based models is that it provides interpretable coefficient estimates that allow us to examine direction and magnitude directly.

The predicted-vs-actual plots for the three final models (LASSO, Ridge, and XGBoost) in Figure 13 confirm the patterns observed in the numerical test results. All models capture the overall positive relationship between the predictors and MATHEFF, as indicated by the upward trend around the 45-degree reference line. However, we can still observe large dispersion of points around the line, which explains the modest R^2 values obtained across all specifications. All

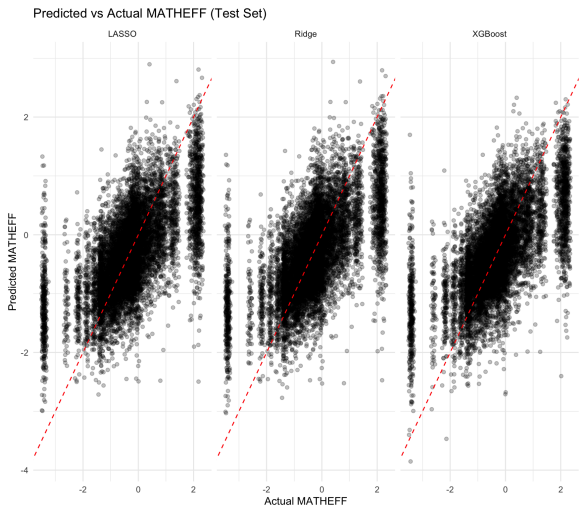

Figure 11: *
Predicted vs. actual MATHEFF

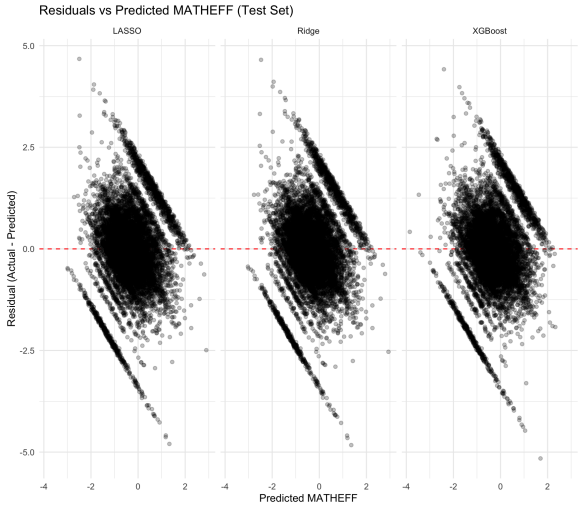

Figure 12: *
Residuals vs. predicted MATHEFF

Figure 13: Predicted vs. actual and residual diagnostics for main models

models perform worse at the extremes of MATHEFF, where predictions become more dispersed, which can reflect the discrete nature of the PISA plausible-value scale and the reduced signal available for extreme-scorers. The residual diagnostic plots for the three main models reinforce these ideas. We can observe that residual variance is lower near the center of the score distribution and substantially larger at the extremes, suggesting that high and low-performing students are more difficult to predict, likely due to higher noise and fewer strong predictors in those ranges. Nevertheless, in both plots XGBoost presents slightly tighter clustering around the diagonal, particularly in the mid-range of the distribution, which aligns with its lower RMSE and higher R^2 on the test set.

Although these results are insightful, there are many limitations that must be addressed. The most important one is computational: the full PISA dataset is too large for my hardware, and repeated memory crashes forced me to work with a reduced sample of 50,000 students with the lowest missingness. This decision kept the analysis feasible but may have introduced subtle selection biases. My imputation strategy is another source of limitation. I initially attempted MICE, but it was computationally impossible at this scale, so I relied on median and mode imputation, which can compress variability and weaken real signals. In addition, by focusing only on conceptual indices, I excluded thousands of item-level variables that may capture nuances lost in the aggregated scales, and the fact that all variables, including MATHEFF, are self-reported means that cultural response styles and subjective biases inevitably shape the data. Finally, limited computational power also constrained hyperparameter tuning, especially for the Random Forest, which likely affected its performance relative to XGBoost.

7 Conclusion

As a brief wrap-up, this project showed that it is possible to capture a non-trivial share of the variation in MATHEFF using machine-learning and penalised linear models. Across methods, motivational, affective, and cognitive-strategy indices consistently emerged as the strongest predictors, while anxiety in mathematics stood out as a robust negative driver. The fact that XGBoost performed best suggests that nonlinear relationships and interactions matter for understanding how students perceive their mathematical capabilities. For future policy work, interventions that reduce mathematics anxiety, strengthen persistence, and build students' familiarity and confidence with mathematical concepts are likely to yield the greatest returns. At the same time, the presence of country effects reinforces that national context and educational systems shape students' beliefs in meaningful ways.

References

Tan, B., & Cutumisu, M. (2022). Employing tree-based algorithms to predict students' self-efficacy in PISA 2018. In A. Mitrovic & N. Bosch (Eds.), *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 634–639). International Educational Data Mining Society. `https://doi.org/10.5281/zenodo.6852970`

Tan, B., Jin, H.-Y., & Cutumisu, M. (2025). Using machine learning algorithms to predict students' general self-efficacy in PISA 2018. *Journal of Applied Developmental Psychology, 99*, 101828.