WE NEED A TITLE HEERE ???

André Costa[1], Celso Ribeiro[1], David Mendes[1]

ISCTE – Instituto Universitário de Lisboa, Portugal

Abstract

This article aims to summarize the approach and results of öz et al. (2024), in which the authors explore employing stacking, boosting, and blending machine learning algorithms to predict student performance in large-scale assessments based on a wide range of predictors and compare their performance.

Keywords: PISA, Machine learning, Stacking, Boosting, Blending, Ensemble learning

1. Introduction

The study öz et al. (2024) uses the PISA 2022 Student questionnaire dataset, focusing on student responses and performance in subjects of mathematics, reading, and science. The authors employ machine learning algorithms to predict student performance, including stacking, blending, and boosting algorithms. The stacking method combines predictions from multiple models to create a meta-model that improves overall prediction performance. Boosting algorithms sequentially train models, each correcting errors from the previous ones, while blending uses a validation set to train the base learners. This study aims to measure the prediction performance of these methods.

1.1. Business Understanding

Large-scale international assessments such as the **Programme for International Student Assessment (PISA)** provide policy-makers with uniquely comparable evidence on the cognitive skills of 15-year-olds. Yet ministries and school leaders frequently lack timely, granular diagnostics that explain *why* some learners underperform while others excel. Our project addresses this gap by developing an AI-based early-warning model that flags students at risk of *low achievement* in mathematics, science, and reading using the publicly released **PISA 2022 Student Questionnaire** micro-data set.

The business—or, more precisely, educational policy—problem can be stated as follows:

> *"Given a fixed budget for instructional support, how can education systems proactively target pedagogical and socio-emotional interventions toward students most likely to obtain low PISA scores, thereby reducing grade repetition and narrowing equity gaps?"*

Operationalising this question involves three goals:

1. **Predictive analytics**: learn a mapping from student-level contextual variables ($\mathbf{X}$) to performance deciles (y) that generalises across participating economies.
2. **Actionable insights**: identify the subset of predictors with the *strongest marginal contribution* to low performance, so interventions can be prioritised.
3. **Benchmarking**: compare the risk profiles of repeating students against high performers, highlighting differences that may travel across national borders.

Key findings to date.. Exploratory correlation and feature-importance screens on $\sim$ 700 candidate variables reveal four robust signals:

- **Socio-economic status (ESCS)** represents the largest single share of variance in low scores; a one-decile drop in ESCS increases the probability of bottom-quartile performance by 6 to 8 pp.
- **Sense of school belonging** and **test-skill confidence** together exert a comparable influence, especially on mathematics.
- **Late school entry** (age > 7) and **grade repetition history** exhibit strong conditional correlations with poor science outcomes.
- Contrary to conventional belief, **time-spent-on-homework** shows only a weak negative correlation once SES and self-efficacy are controlled.

These empirical patterns motivate our choice of ensemble models (stacking, gradient boosting) that handle high-dimensional, collinear, and partially missing data while producing interpretable feature importances.

Why cross-country comparison matters.. Although PISA is standardised, the institutional context in which 15-year-olds learn varies considerably:

Recognising such systemic contrasts is essential: a "risk factor" identified in one jurisdiction (e.g. after-school tutoring hours in Shanghai) may not be policy-controllable in another (e.g. England, where after-school tuition is privately financed).

Therefore, our modelling pipeline includes country-specific fixed effects and studies interaction terms between student attributes and system-level dummies.

What teaching strategies enhance reading performance?. Insights from **PISA 2018** contextual data offer valuable evidence. According to the teacher questionnaire, strategies such as *explicit reading instruction*, *activating prior knowledge*, and *engaging students in metacognitive practices* were correlated with higher reading performance. From the student perspective, respondents who reported frequent use of *discussions about texts*, *clarity in lesson structure*, and *constructive feedback* also scored higher on average. Interestingly, while both teachers and students pointed to structured and cognitively activating instruction as beneficial, teachers tended to emphasize planning and scaffolding, whereas students highlighted motivation and classroom climate. This discrepancy reinforces the importance of triangulating perspectives in education data analytics.

Profiling students in vocational education tracks.. Another avenue of investigation focuses on students enrolled in **vocational and professional training programs** as defined in PISA. Historically, these students perform lower in mathematics than those in general education tracks. Variables explaining this gap include reduced parental education, lower socio-economic status, and limited exposure to advanced mathematics content. However, over successive cycles of PISA (2012–2018), some countries have seen marginal gains for vocational-track students—particularly where applied mathematics and contextualised problem-solving were integrated into the curriculum. Our model incorporates program type as a categorical feature and enables interaction testing with SES, gender, and country-level effects to better understand these evolving patterns.

Success criteria.. From a business-value perspective, we deem the CRISP-DM *Business Understanding* phase complete when we can:

- articulate the decision points (resource targeting, curriculum design) that model outputs will inform;
- translate model metrics into cost–benefit terms for ministries (e.g. reduction in misallocated tutoring hours per $100k invested in data collection);
- outline ethical safeguards to prevent algorithmic bias against vulnerable groups.

Establishing this shared understanding with stakeholders ensures that the subsequent CRISP-DM phases—Data Understanding, Data Preparation, Modelling, and Evaluation—remain anchored to actionable educational impact.

Building on this policy motivation, the empirical backbone of our work is the study by öz et al. (2024), which exploits the same **PISA 2022 Student Questionnaire** micro-data. Their benchmark investigation compares three ensemble-learning families—*stacking*, *boosting*, and *blending*—for predicting mathematics, reading, and science scores. Stacking aggregates the outputs of diverse base learners into a meta-model, boosting trains learners sequentially so each iteration corrects the residuals of the previous one, while blending fits base models on a hold-out validation set before meta-aggregation. öz et al. report that stacking consistently yields the lowest error metrics (MAPE, MAE, MSE), a finding that motivates our own choice of stacked architectures for early-warning classification.

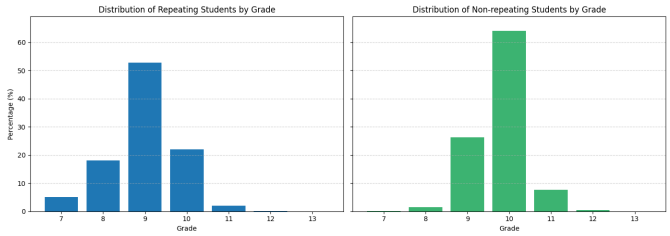

Figure 1: Distribution of students by grade repetition status.

2. Data Understanding

2.1. *Question 1 – Profile of repeating students in PISA. Which variables contribute to explaining their performance in mathematics? How has it evolved over the different assessment cycles?*

For this question, we used the Student Questionnaire data from 2022. (*citar fonte*).

This dataset contains information about the students' backgrounds, their attitudes towards mathematics, and their performance in the PISA assessment. It has 1278 columns, from which 1260 are numerical and 18 are categorical. It has 613744 rows, where each row corresponds to a student. Around 10% of this students are repeating students, which means they have been retained in the same grade for at least one year. In Figure 1 we can see the distribution of students by grade repetition status.

The target variable for this analysis is the student's performance in mathematics. Their score can be calculated as the average of the values across all "Possible Math Value" columns, which are represented in the dataset from `PV1MATH` to `PV10MATH`. These features are plausible values, each representing multiple estimates of the student's performance. Averaging them provides a more reliable and comprehensive measure of the student's grade. This score is a continuous variable, and can vary from a minimun of 0 and a maximum of 1000.

We then conducted an exploratory analysis of the dataset to identify relevant patterns and potential outliers. This helped us uncover special cases that may influence student performance, such as differences in educational systems across countries. For instance, the English school system has a different structure compared to the rest of the world (*citar fonte*), which may affect the results. In Figure 2, we can see that most of the students in 11th, 12th and 13th grades are from the UK. This difference may lead to a disproportionate representation of students and we should consider them as an exception in the next phases.

For the categorical variables

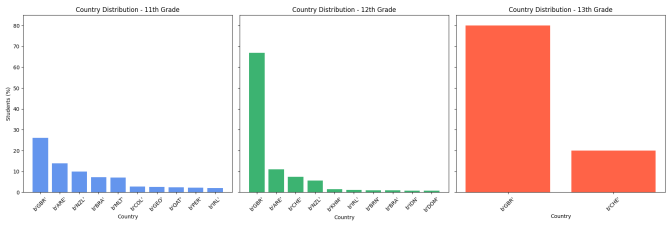

Figure 2: Distribution of students by country (%).

2.2. *Question 2*

2.3. *Question 3*

3. Data Preparation

3.1. *Question 1*

4. Methods

4.1. *Stacking*

Stacking is an ensemble learning technique developed by Wolpert (1992) that combines several prediction models and uses their outputs as input for a next-level model (meta-model), aiming to improve overall prediction performanceöz et al. (2024). In the study öz et al. (2024), the authors divide stacking into level 0 and level 1, where layer 0 refers to where the different models generate distinct predictions, and level 1 combines predictions of level 0. For level 0, some of the different models that could be used are DTs, NNs, SVMs, and kNN, in which the resulting predictions are then used and combined in layer 1, in order to improve the overall prediction Wolpert (1992), the authors choose to use ridge regression, which can be useful to avoid overfitting Cui et al. (2021)

4.2. *Boosting*

Because the dataset used in this study is composed of missing data, it is imperative to implement learners that deal with missing data, the boosting algorithms used were XGBoost, HGB, and LightGBM öz et al. (2024)

4.3. *blending*

Blending and stacking are both techniques used in ensemble learning that share similarities. The key difference between them is that in blending, the base learners are trained using predictions obtained from the validation set instead of directly from the training setöz et al. (2024)

5. Results

The study found that the stacking method produced the lowest error metrics (Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE)) for most countries compared to boosting and blending. Robust linear mixed-effects models indicated that stacking had significantly lower error metrics across all subjects as showed in 3.

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

Figure 3: The Number (%) of the countries exhibiting the lowest error values generated by each algorithm for all subjects öz et al. (2024)

6. Conclusion

The study öz et al. (2024) concludes that stacking is a favorable option for better performance, stable and accurate predictions based on evaluation with various metrics such as MAPE, MAE, MSE and RMSE where it obtained a significant metric score in all subjects compared to boosting and blending. Due to this study only evaluating the PISA dataset, I think it would be interesting to evaluate it with other datasets to improve this analysis further

References

Cui, S., Yin, Y., Wang, D., Li, Z., Wang, Y., 2021. A stacking-based ensemble learning method for earthquake casualty prediction. Applied Soft Computing 101, 107038. URL: https://www.sciencedirect.com/science/article/pii/S156 doi:https://doi.org/10.1016/j.asoc.2020.107038.

Wolpert, D.H., 1992. Stacked generalization. Neural Networks 5, 241–259. URL: https://www.sciencedirect.com/science/article/pii/S089 doi:https://doi.org/10.1016/S0893-6080(05)80023-1.

öz, E., Bulut, O., Cellat, Z.F., Yürekli, H., 2024. Stacking: An ensemble learning approach to predict student performance in pisa 2022. Education and Information Technologies , 1–27doi:10.1007/s10639-024-13110-2.