International Journal of Assessment Tools in Education
2026, Vol. 13, No. 3, 957–977
https://doi.org/10.21449/ijate.1856645
Journal homepage: https://dergipark.org.tr/en/pub/ijate *Research Article*

Benchmarking and interpreting a tabular foundation model for PISA 2022 mathematical literacy

Büşra Eren [1*], **Tuğba Niksarlı** [2]

[1]National Defence University, Army NCO Vocational School, Department of Educational Sciences, Balıkesir, Türkiye
[2]National Defence University, Army NCO Vocational School, Balıkesir, Türkiye

ARTICLE HISTORY

Received: Jan. 05, 2026

Accepted: June 03, 2026

KEYWORDS

Large-scale assessment,
Machine learning,
TabPFN,
Learning analytics,
Mathematical literacy.

Abstract: Recent advances in machine learning have expanded the analytical toolkit available for large-scale assessment data; however, many high-performing approaches remain limited in terms of interpretability and their suitability for assessment-informed decision-making. This study used data from 7,250 students in the Türkiye sample of PISA 2022 mathematical literacy to benchmark and interpret the Tabular Prior-Data Fitted Network (TabPFN) for classifying student achievement. For the purpose of model evaluation, mathematical literacy was operationalised as the mean of ten plausible values and then converted into a binary outcome using the PISA Level 2 proficiency benchmark, consistent with the study's classification-oriented aim. TabPFN was compared with eXtreme Gradient Boosting (XGBoost), Light Gradient-Boosting Machine (LightGBM), Stacking, Random Forest, and Logistic Regression under consistent preprocessing and stratified cross-validation procedures. Beyond predictive performance, feature-importance patterns were examined using Shapley Additive Explanations (SHAP) for TabPFN and permutation importance for the baseline models. The findings showed that TabPFN achieved competitive performance and produced an interpretable predictor profile. Home resources and mathematics self-efficacy emerged as the most influential predictors, followed by curiosity and classroom disciplinary climate. These results suggest that tabular foundation models can support large-scale educational assessment not only by providing accurate classifications, but also by generating interpretable evidence about achievement-related learner and contextual factors. The findings should be interpreted as classification evidence around a meaningful proficiency threshold rather than as a model of continuous mathematics achievement.

INTRODUCTION

In contemporary educational evaluation and student assessment, the proliferation of digital record systems, large-scale assessments, and institutional data infrastructures has increasingly positioned machine learning (ML) as a central analytic approach for making sense of complex information about students and schools. Compared with conventional parametric models, ML techniques can accommodate nonlinear relationships and higher-order interactions in complex feature spaces, often yielding stronger predictive performance and more detailed portraits of

*Corresponding author: Büşra EREN busra.eren@msu.edu.tr National Defence University, Army NCO Vocational School, Department of Educational Sciences, Balıkesir, Türkiye

e-ISSN: 2148-7456

learner behaviour (Baker & Inventado, 2014; Holmes *et al.*, 2019; Kizilcec, 2016; Romero & Ventura, 2020).

In this study, artificial intelligence (AI) is conceptualised as the broad umbrella term for computational systems that support data-informed reasoning, prediction, and decision-making in educational contexts. ML is defined more specifically as data-driven algorithms that learn predictive patterns from observed data, whereas deep learning is understood as a subset of machine learning based on multilayer neural architectures. This distinction is important because artificial intelligence in general or deep learning models designed for unstructured data are not evaluated in the present study; rather, the focus is placed on supervised machine learning models for tabular large-scale assessment data, with Tabular Prior-Data Fitted Network (TabPFN) positioned as a recent tabular foundation model within this modelling landscape (Hollmann *et al.*, 2023; Holmes *et al.*, 2019; Hospedales *et al.*, 2022; Mislevy, 2018).

The central research problem addressed in this study is that, although machine learning models are increasingly used to predict academic outcomes in large-scale educational data, their contribution to student assessment remains limited when predictive performance is examined separately from interpretability and measurement-relevant decision use. In research based on the Programme for International Student Assessment (PISA), this issue is especially important because student achievement is estimated through plausible values and accompanied by rich contextual questionnaire indicators, requiring models to be evaluated not only for classification accuracy but also for whether their predictions are linked to educationally meaningful learner and contextual variables (OECD, 2023; Rutkowski *et al.*, 2010). Accordingly, model performance and interpretability are treated in the present study as connected aspects of assessment-informed modelling rather than as separate technical objectives.

A wide range of algorithms has been applied in this space, including classical supervised learners, tree-based ensembles, stacking approaches, and more recent foundation-model-based methods. In the present study, the selected models were intended to represent complementary families of supervised learning approaches commonly used with tabular educational data. Logistic regression was included as a transparent linear baseline, random forest as a bagging-based tree ensemble, eXtreme Gradient Boosting (XGBoost) and Light Gradient-Boosting Machine (LightGBM) as gradient-boosted tree models, stacking as an ensemble strategy that combines multiple learners, and TabPFN as a recent tabular foundation model that offers single-step prediction with minimal task-specific tuning. Through this model set, a spectrum of approaches ranging from interpretable conventional classification to high-capacity ensembles and foundation-model-based prediction could be compared (Breiman, 2001; Chen & Guestrin, 2016; Hollmann *et al.*, 2023; Hollmann *et al.*, 2025; Öz *et al.*, 2025; Romero & Ventura, 2020; Slater *et al.*, 2017; Vanschoren, 2019).

From the perspective of educational evaluation, however, two limitations recur. First, empirical studies often foreground global performance metrics while offering limited discussion of how model outputs can be translated into concrete assessment practices, such as the early identification of students who are struggling or the targeted allocation of support in authentic institutional settings (Romero & Ventura, 2020; Slater *et al.*, 2017). Second, interpretability is addressed unevenly: comparatively few contributions have systematically examined the internal behavior of the models to identify stable, theory-aligned determinants of achievement that educators and researchers can meaningfully use in student assessment and decision-making (Lundberg & Lee, 2017; Molnar, 2025).

International large-scale assessments provide a promising, yet still underused, context for tackling these shortcomings. PISA, in particular, combines standardised achievement measures with extensive student- and school-level background questionnaires, enabling multilevel explanatory and predictive analyses that connect performance to rich contextual information (Akyüz & Pala, 2010; Kamaliyah *et al.*, 2013; OECD, 2023). In the 2022 cycle, mathematical

literacy was the major assessment domain; therefore, the focus of the present study was placed on PISA 2022 mathematical literacy rather than on the entire PISA assessment framework. Against this backdrop, models such as TabPFN have emerged as approaches that seek to streamline modelling workflows while retaining competitive performance. These models enable single-step prediction without requiring manual model selection or extensive hyperparameter tuning, thereby contributing to reductions in computational burden, expertise demands, and time requirements in educational measurement systems (Hollmann *et al.*, 2023; Hollmann *et al.*, 2025; Vanschoren, 2019).

The analytic structure of the present study reflects this design. The dataset consists of student-level PISA 2022 Türkiye data, including ten plausible values for mathematical literacy and selected student background-questionnaire indices representing home resources, parental education, motivational and affective characteristics, classroom climate, instructional experiences, and family or peer support. For the classification task, the continuous mathematical literacy indicator was derived by averaging the ten plausible values and was then operationalised as a binary outcome using the PISA Level 2 proficiency benchmark. This structure made it possible to examine both how accurately different models classified students around a meaningful proficiency threshold and which learner- and context-related indicators contributed most strongly to these classifications (OECD, 2023; Rutkowski *et al.*, 2010).

Post-hoc interpretability techniques—most notably Shapley Additive Explanations (SHAP)—offer a practical mechanism for narrowing the gap between predictive performance and actionable educational insight. Existing reviews, however, point to the inconsistent use of such techniques and to the limited number of studies that combine competitive baseline models with transparent explanations on PISA 2022 mathematical literacy data (Peña-Ayala, 2014; Romero & Ventura, 2020).

In this evaluation study, the performance of TabPFN on PISA 2022 mathematical literacy data from Türkiye was examined, and the model was situated within the context of student assessment and educationally meaningful interpretation. The study is organised around two interrelated components: benchmarking TabPFN against widely used classification models and interpreting the predictor-importance patterns underlying achievement classification. Within this structure, benchmarking corresponds to RQ1 and RQ2, whereas interpretation corresponds to RQ3. Three main objectives were addressed: (a) to evaluate the extent to which TabPFN effectively classifies mathematics achievement in the Türkiye sample; (b) to compare its performance with widely used models, including random forest, logistic regression, stacking, LightGBM, and XGBoost; and (c) to identify the most influential predictors of mathematical literacy by using SHAP for TabPFN and permutation importance for the remaining models.

Guided by these aims, the following research questions were addressed:

RQ1. To what extent does TabPFN accurately classify students at or above and below the PISA Level 2 proficiency benchmark in PISA 2022 Mathematical Literacy?

RQ2. How does TabPFN's classification performance compare with that of XGBoost, LightGBM, stacking, logistic regression, and random forest?

RQ3. Which variables contribute most strongly to the prediction of mathematical literacy?

By integrating model benchmarking with predictor-level interpretation, this study is intended to bridge methodological innovation and educational evaluation. Accordingly, the study contributes to the literature by integrating comparative model evaluation with interpretable predictor evidence in the context of PISA 2022 mathematical literacy.

METHOD

Study Design

This study employed a secondary data analysis and model-evaluation design using the Türkiye sample of PISA 2022, coordinated by the Organisation for Economic Co-operation and

Development (OECD). The analysis was framed around predictive model evaluation for student assessment. Because mathematics was the major domain in PISA 2022, the study was restricted to mathematical literacy outcomes and domain-relevant student-level predictors. The primary aim was to evaluate TabPFN for classifying students' mathematics achievement in a large-scale assessment context and to benchmark its performance against XGBoost, LightGBM, random forest (RF), logistic regression (LR), and a stacking ensemble commonly used in educational data-mining research (Öz *et al.*, 2025; Slater *et al.*, 2017).

Because the primary objective was algorithmic benchmarking rather than population parameter estimation, PISA sampling weights were not incorporated in model training or evaluation. Accordingly, the findings should be interpreted as evidence of comparative model performance within the analysed sample, rather than as weighted population estimates for Türkiye.

Participants and Data Source

The dataset was drawn from the Türkiye sample of PISA 2022, comprising 7,250 15-year-old students who completed both the cognitive assessments and the background questionnaires. PISA 2022 was coordinated by the OECD and implemented through a stratified sampling design and standardised administration procedures, designed to support nationally representative estimates of student achievement and contextual indicators (OECD, 2009; OECD, 2023). The analytic focus was limited to mathematical literacy because this domain constituted the major assessment area of the 2022 cycle. The present study relied exclusively on de-identified public-use data, and no additional data collection was undertaken by the authors.

Outcome Variable

Mathematical literacy formed the basis of the classification outcome and was represented in PISA 2022 by ten plausible values (PV1MATH–PV10MATH), which provide a more robust representation of student proficiency. For model evaluation, the mean of these ten plausible values was used to derive a single student-level composite indicator of mathematics proficiency. This continuous indicator was then converted into a binary outcome in line with the study's classification-oriented research questions, model-comparison design, and interpretability focus. Students were classified with reference to the achievement threshold proposed by Hanushek and Woessmann (2011), which operationally corresponds to the widely used PISA Level 2 proficiency benchmark.

In OECD reporting, Level 2 is generally regarded as a baseline proficiency level, indicating that students have begun to demonstrate the core mathematical competencies required to interpret and use mathematics in real-life and relatively structured problem situations (OECD, 2023). Accordingly, students whose mean PV1MATH–PV10MATH score fell below this benchmark were coded as unsuccessful, whereas those meeting or exceeding the benchmark were classified as successful. The resulting binary outcome showed a mildly imbalanced distribution between successful and unsuccessful students; therefore, this class distribution was considered when reporting and interpreting model performance.

Predictors and Data Collection Instruments

The predictor set was constructed using student-level PISA 2022 indices that capture key dimensions commonly examined in research on mathematics achievement and educational equity, including student background, home resources, classroom climate, and learner dispositions (Akyüz & Pala, 2010; OECD, 2023; Rutkowski *et al.*, 2010). In line with the study's focus on interpretable student assessment, attention was restricted to a compact set of theoretically grounded covariates that can plausibly support the interpretation of student achievement classifications. All predictors were derived from PISA's standardised cognitive assessments and student background questionnaires.

Predictor Selection Rationale. The selection of predictors was guided by three complementary considerations: alignment with the PISA 2022 mathematics and background-questionnaire framework, empirical relevance in prior PISA-based mathematics achievement research, and the need for a compact and interpretable feature set for model comparison and feature-importance analysis. The PISA 2022 framework links student achievement with contextual information collected through student and school questionnaires, including home resources, family background, learning-related beliefs, classroom experiences, and broader learning conditions (OECD, 2023).

Accordingly, the retained predictors were organised into four substantive groups. The first group represented socioeconomic and home-resource conditions, including HOMEPOS, FISCED, and MISCED. The second group represented mathematics-related motivational and affective characteristics, including MATHEFF, ANXMAT, MATHPERS, PERSEVAGR, and CURIOAGR. The third group represented classroom and instructional-environment indicators, including DISCLIM, TEACHSUP, and COGACRCO. The fourth group represented family or peer-support conditions, including FAMSUP and COOPAGR.

A recent systematic review of PISA-based mathematics achievement studies reported that mathematics self-efficacy, mathematics anxiety, home educational resources, parental education, classroom climate, and learning opportunities are among the most frequently examined predictors of mathematics performance (Wang *et al.*, 2023). Earlier cross-national PISA evidence similarly showed that socioeconomic status, mathematics self-efficacy, and mathematics anxiety are associated with mathematics achievement across several countries, including Türkiye (Bakan Kalaycıoğlu, 2015). In addition, disciplinary classroom climate and mathematics self-efficacy have been linked to mathematics performance in PISA-based analyses (Cheema & Kitsantas, 2014), while recent PISA 2012–2022 evidence highlights the relevance of instructional-quality indicators such as teacher support, disciplinary climate, and cognitive activation when socioeconomic background is taken into account (Liu *et al.*, 2024).

Finally, the predictor set was kept compact to support interpretability and to minimise conceptual overlap and content redundancy among questionnaire-based indicators. Because the purpose of the study was not to maximise prediction through an unrestricted feature pool, but to compare model performance and examine stable predictor-importance patterns, highly overlapping or overly broad questionnaire constructs were avoided when more interpretable indicators were available.

The retained PISA variable names were GENDER, HOMEPOS, FISCED, MISCED, DISCLIM, TEACHSUP, COGACRCO, MATHEFF, PERSEVAGR, COOPAGR, MATHPERS, ANXMAT, CURIOAGR, and FAMSUP. Because these official variable names are used in the Findings section and in the feature-importance tables, they were retained consistently throughout the manuscript. For clarity, their meanings are stated once here: GENDER = student gender; HOMEPOS = home possessions; FISCED = father's highest ISCED level of education; MISCED = mother's highest ISCED level of education; DISCLIM = disciplinary climate in mathematics lessons; TEACHSUP = mathematics teacher support; COGACRCO = cognitive activation in mathematics; MATHEFF = mathematics self-efficacy; PERSEVAGR = general perseverance; COOPAGR = cooperation; MATHPERS = mathematics perseverance; ANXMAT = mathematics anxiety; CURIOAGR = curiosity; and FAMSUP = family support.

Preprocessing and Data Preparation

Because TabPFN operates on fully observed numeric tables (Hollmann *et al.*, 2023), a preprocessing procedure aligned with the cross-validation framework was implemented to reduce information leakage (Witten *et al.*, 2016). Before preprocessing, five unusable student records were excluded from the initial downloaded dataset because they had missing values across all variables included in the analysis. After this exclusion, the final analytic sample

consisted of 7,250 students. Before imputation, the modelling dataset consisted of 7,250 students, 14 predictors, and the binary outcome. Missing values were observed only in the predictor variables and remained below 5% for each predictor. Therefore, no additional student records were removed from the sample due to missing predictor data, and the sample size was retained at N = 7,250. Missing values in continuous predictors were handled using median imputation, whereas categorical predictors were completed with the modal category. Median imputation was preferred as a robust procedure because it is less sensitive to extreme values and distributional asymmetry than mean imputation, which is particularly relevant for questionnaire-based indices. In each cross-validation fold, imputation parameters were estimated only from the training subset and then applied unchanged to the corresponding validation and test subsets.

Pearson correlation analysis and collinearity diagnostics were also conducted as preliminary data-preparation steps to identify variables with weak predictive relevance or redundant information; these checks were used for data preparation rather than inferential interpretation (Hosmer *et al.*, 2013). All predictor-evaluation steps related to complete missingness, near-zero variance, collinearity problems, and weak association with the binary outcome were conducted within each cross-validation fold using only the training subset. Conducting this process separately within each cross-validation fold ensured that decisions regarding redundant information, collinearity problems, and variable exclusion were made without being influenced by the validation or test data. Categorical predictors were encoded in each fold using information obtained from the training data. Tree-based learners, including XGBoost, LightGBM, RF, and the base learners in the stacking ensemble, were trained on unstandardised inputs, whereas standardisation was applied to the predictors used by logistic regression and the stacking meta-learner.

Following these steps, the modelling dataset was transformed into a fully observed and entirely numeric structure, allowing TabPFN to be applied directly without task-specific hyperparameter tuning (Hollmann *et al.*, 2023). All preprocessing operations were defined within the cross-validation loop rather than on the full dataset. This approach supported replicability and ensured consistency with recommended practices for model evaluation in educational data mining (Witten *et al.*, 2016).

Machine Learning Models

Building on the predictor set and the leakage-safe, fold-wise preprocessing described above, six complementary models that are routinely used in educational data mining and that span a range of complexity and interpretability were compared. All models were estimated within the same cross-validation framework to provide a fair, decision-relevant benchmark for student assessment on PISA 2022 Mathematical Literacy data.

TabPFN

TabPFN is a transformer-based meta-learning model trained on a large prior of simulated tabular tasks. TabPFN is designed to deliver high-quality, single-step predictions without requiring manual model selection or extensive hyperparameter tuning, thereby reducing the level of expertise and computational resources typically needed to approach state-of-the-art performance (Hollmann *et al.*, 2023, 2025; Hospedales *et al.*, 2022; Vanschoren, 2019).

XGBoost and LightGBM

XGBoost and LightGBM (gradient-boosted trees) represent widely adopted implementations of gradient-boosted decision trees and are commonly used as strong baseline models for tabular prediction tasks, as they combine competitive predictive accuracy with scalability to large and complex datasets (Chen & Guestrin, 2016; Öz *et al.*, 2025).

Random forest

Random forest constructs an ensemble of decision trees grown on bootstrapped samples and random subsets of predictors, aggregating their outputs to yield robust performance across a variety of conditions (Biau & Scornet, 2016; Breiman, 2001; Seni & Elder, 2010).

Logistic regression

Logistic regression is a transparent probabilistic classifier with interpretable coefficients and odds ratios and remains a common baseline in educational research (Hosmer *et al.*, 2013; Menard, 2002). In this study, standard (unpenalised) LR with main effects only was fitted. Neither quasi-separation nor rare-event problems were detected, and formal regularization was therefore not required.

Stacking

In the stacking ensemble, level-0 base learners (XGBoost, LightGBM, RF, and LR) generated out-of-fold (OOF) predictions, which were then used as inputs to a logistic regression meta-learner (Sill *et al.*, 2009; Ting & Witten, 1999; van der Laan *et al.*, 2007; Wolpert, 1992). Restricting the meta-learner to OOF predictions was intended to exploit complementary error profiles among the base models while limiting overfitting and preserving generalisation performance in student assessment applications.

Taken together, these models reflect current practice in educational data mining and cover a spectrum that balances predictive strength and interpretability, from fully transparent LR to high-capacity ensembles and the tabular foundation model TabPFN (Öz *et al.*, 2025; Romero & Ventura, 2020; Slater *et al.*, 2017).

Model Training and Evaluation

All models were trained and evaluated on the binary success outcome using stratified k-fold cross-validation. Within each fold, the preprocessing steps described above were fitted on the training subset and then applied unchanged to the corresponding validation and test subsets, after which each model produced predicted probabilities for student success. Predicted probabilities were converted into class labels using a fixed decision threshold of 0.50 for all models. This common threshold was used to ensure a consistent comparison across classifiers; no model-specific threshold optimisation was performed. Because threshold-dependent metrics may vary according to the selected decision rule, accuracy and F1-score were interpreted alongside threshold-independent discrimination, primarily ROC–AUC, following recommendations to evaluate classification models using multiple complementary performance indicators rather than relying on a single metric (Sokolova & Lapalme, 2009). Performance was summarised using receiver operating characteristic area under the curve (ROC–AUC), accuracy, and F1-score, with the mildly imbalanced class distribution taken into account when interpreting these metrics. Fold-wise results were averaged to obtain overall estimates for each model, thereby providing a fair comparison between TabPFN and the ensemble and regression baselines in the context of large-scale student assessment.

Software and Computational Environment

The analyses were carried out in Python. Data management and preprocessing steps were implemented with pandas and NumPy (Harris *et al.*, 2020), whereas model fitting, cross-validation, and performance evaluation were mainly conducted through scikit-learn (Pedregosa *et al.*, 2011). Model-specific packages were additionally used for XGBoost, LightGBM, TabPFN, and SHAP-based interpretation, in line with the corresponding methodological literature (Chen & Guestrin, 2016; Hollmann *et al.*, 2023; Ke *et al.*, 2017; Lundberg & Lee, 2017). Fold-level ROC–AUC comparisons were performed using Python-based statistical procedures. To enhance reproducibility, all models were evaluated under the same cross-validation design, preprocessing pipeline, and fixed classification threshold.

RESULTS

In relation to RQ1, TabPFN was first compared with commonly used baselines in classifying students' mathematics achievement. Because mathematical literacy was operationalised as a binary outcome around the PISA Level 2 proficiency benchmark, the findings should be interpreted as evidence about classification performance near a meaningful proficiency threshold rather than as estimates of continuous achievement. This framing directly corresponds to RQ1 and RQ2, which focus on TabPFN's classification performance and its comparative standing against baseline models. All models were evaluated using stratified (class-proportion–preserving) 10-fold cross-validation, and fold-wise summary statistics ($M \pm SD$) are reported in Table 1. Additional fold-wise ROC–AUC comparison (Figure S1), model-specific ROC curves (Figure S2), confusion matrices (Table S1), and feature-importance plots (Figure S3) are provided in Appendix A.

Table 1. *Comparative performance of machine learning models based on 10-fold cross-validation.*

Model	ROC-AUC (±*SD*)	Accuracy (±*SD*)	Precision (±*SD*)	Recall (±*SD*)	F1 (±*SD*)
LR	.792 ± .019	.727 ± .020	.719 ± .014	.693 ± .013	.706 ± .013
XGBoost	.812 ± .020	.738 ± .021	.730 ± .018	.708 ± .017	.719 ± .017
LightGBM	.805 ± .020	.728 ± .019	.721 ± .021	.694 ± .016	.707 ± .015
RF	.812 ± .017	.737 ± .012	.736 ± .019	.693 ± .017	.713 ± .015
Stacking	.816 ± .019	.742 ± .015	.737 ± .018	.708 ± .016	.722 ± .014
TabPFN	.818 ± .019	.743 ± .020	.737 ± .017	.710 ± .015	.723 ± .010

As summarised in Table 1, TabPFN obtained the strongest overall results across the primary indicators (ROC–AUC = .818; accuracy = .743; precision = .737; recall = .710; F1 = .723). The remaining metrics followed the same broad ordering, with stacking and XGBoost closest to TabPFN. These patterns indicate that TabPFN provided the strongest descriptive performance, but the magnitude of the differences suggests an incremental rather than a decisive advantage over the strongest ensemble baselines. The relatively close performance of TabPFN, stacking, XGBoost, and RF suggests that several high-capacity models captured similar nonlinear and interaction-based information in the selected PISA predictors, whereas the lower performance of LR is consistent with its more restrictive linear specification. These descriptive patterns were examined formally using statistical tests (Table 2). Because the primary focus was on threshold-independent discrimination, probabilistic calibration indices were not additionally examined. To assess whether the observed gaps in ROC–AUC were statistically reliable, fold-level AUC values were analysed using a Friedman framework with Holm-adjusted pairwise contrasts. The resulting mean differences, effect sizes, and p values are reported in Table 2.

Table 2 reports the pairwise differences in fold-level ROC–AUC, together with Holm-adjusted p values and Cohen's d effect sizes, and broadly confirms the descriptive ordering observed in Table 1. TabPFN achieved significantly higher ROC–AUC than LR (ΔAUC = –.026, p = .001), LightGBM (ΔAUC = –.012, p = .001), and RF (ΔAUC = –.006, p = .019). Although these absolute AUC gaps are modest, the corresponding effect sizes indicate marked improvements over LR and LightGBM and a clear advantage over RF. In contrast, the TabPFN–XGBoost comparison yielded a small, non-significant difference (ΔAUC ≈ .006, p = .083), and the contrast between TabPFN and stacking was also small and non-significant (ΔAUC ≈ –.001, p = .556). Taken together, the practical ordering can be summarised as TabPFN ≈ stacking ≥ XGBoost ≈ RF ≥ LightGBM ≥ LR. Thus, with respect to RQ2, the inferential results support the conclusion that TabPFN outperformed LR, LightGBM, and RF on fold-level ROC–AUC, but did not show a statistically reliable advantage over stacking or XGBoost. Accordingly, the comparative finding is best interpreted as competitive performance and near-parity with the strongest baselines rather than uniform superiority over all models.

Table 2. *Friedman test results for ROC–AUC comparisons (10-fold cross-validation).*

Model comparison	Mean difference	Cohen's *d*	*p*
LR vs XGBoost	–.020	–1.687	.001
LR vs TabPFN	–.026	–3.791	.001
LR vs Stacking	–.024	–2.780	.001
LightGBM vs XGBoost	–.006	–2.222	.001
LR vs RF	–.019	–1.230	.001
LightGBM vs Stacking	–.011	–2.431	.001
LightGBM vs TabPFN	–.012	–1.462	.001
LightGBM vs LR	.013	1.085	.005
LightGBM vs RF	–.005	–0.896	.009
Stacking vs XGBoost	.004	1.078	.019
RF vs TabPFN	–.006	–1.030	.019
RF vs Stacking	–.004	–0.133	.019
TabPFN vs XGBoost	.005	–1.083	.083
Stacking vs TabPFN	–.001	–0.254	.556
RF vs XGBoost	–.000	–0.072	.492

Note. Holm-adjusted *p* values.

Having established this comparative ordering, the analysis turned to RQ3 and examined the relative contributions of individual predictors. To maintain the logical sequence of the research questions, the analysis now shifts from model-level benchmarking evidence for RQ1–RQ2 to predictor-level interpretation for RQ3. For the five baseline models, predictor influence was quantified using permutation importance; for TabPFN, mean absolute SHAP values computed exclusively from out-of-fold (OOF) predictions were used to preserve comparability and avoid information leakage.

Table 3. *Feature-importance magnitudes by model.*

	Model					
Variable	LightGBM	LR	RF	Stacking	XGBoost	TabPFN
MATHEFF	.095	.091	.095	.094	.090	.093
HOMEPOS	.086	.107	.092	.096	.085	.072
CURIOAGR	.017	.016	.014	.016	.018	.015
DISCLIM	.014	.011	.013	.013	.012	.013
FISCED	.010	.013	.008	.009	.010	.009
TEACHSUP	.000	.000	.000	.000	.000	.006
FAMSUP	.007	.000	.003	.004	.008	.006
COGACRCO	.000	.000	.000	.000	.005	.006
COOPAGR	.006	.003	.003	.004	.008	.005
MISCED	.006	.010	.004	.006	.008	.004
ANXMAT	.000	.000	.000	.000	.000	.004
MATHPERS	.006	.001	.005	.004	.000	.002
GENDER	.006	.000	.003	.003	.005	.002
PERSEVAGR	.000	.000	.000	.000	.000	.001

Note. For the five baseline models, feature importance reflects the mean decrease in ROC–AUC when the corresponding predictor is randomly permuted. For TabPFN, importance values are the mean absolute SHAP scores computed from out-of-fold predictions.

Numerical summaries are reported in Table 3 (importance magnitudes) and Table 4 (rankings), with model-specific panels provided in Appendix A (Figure S3). Because permutation importance and SHAP differ in their formal definitions and operate on different scales, absolute

magnitudes are not interpreted across models; instead, the analysis focused on within-model rank patterns and convergences in ranking across learners (Lundberg & Lee, 2017; Molnar, 2025). Cross-model rank concordance, quantified via Spearman's ρ (Table 4), corroborated the patterns observed in the importance magnitudes.

Table 4. *Feature-importance rankings by model (permutation/SHAP; stratified 10-fold cross-validation, highest to lowest).*

	Model						
Variable	LightGBM	LR	RF	Stacking	XGBoost	TabPFN	Aggregate ranking
MATHEFF	1	2	1	2	1	1	1
HOMEPOS	2	1	2	1	2	2	2
CURIOAGR	3	3	3	3	3	3	3
DISCLIM	4	5	4	4	4	4	4
FISCED	5	4	5	5	5	5	5
MISCED	8	6	7	6	8	10	6
FAMSUP	6	10	8	9	6	7	7
COOPAGR	10	7	9	8	7	9	8
MATHPERS	9	8	6	7	11	12	9
GENDER	7	9	10	10	10	13	10
TEACHSUP	11	11	11	11	11	6	11
COGACRCO	11	11	11	11	9	8	11
ANXMAT	11	11	11	11	11	11	12
PERSEVAGR	11	11	11	11	11	14	13

Inspection of Table 4 shows a highly consistent pattern across models. HOMEPOS and MATHEFF emerge as the dominant predictors, occupying the first or second position in every learner. In most cases, this top layer is followed by CURIOAGR and DISCLIM, which form a recurrent upper-middle tier. FISCED and MISCED, together with FAMSUP and COOPAGR, generally make mid-range contributions, whereas MATHPERS and GENDER tend to appear in lower positions. TEACHSUP, COGACRCO, ANXMAT, and PERSEVAGR show consistently low importance values across models. With respect to RQ3, this convergence across model families indicates that the feature-importance results were not driven by a single algorithmic specification. Instead, the same broad pattern appeared across linear, tree-based, ensemble, and foundation-model-based approaches, strengthening the interpretation that home resources, mathematics self-efficacy, curiosity, and disciplinary climate were the most stable predictors of the binary mathematical literacy classification.

This configuration is visible both in the importance magnitudes (Table 3) and in the model-specific panels (Appendix A, Figure S3) and is summarized in the aggregate ranking in Table 4. Taken together, the findings suggest that, irrespective of model family, (a) HOMEPOS and MATHEFF systematically drive the classification signal, (b) CURIOAGR and DISCLIM form a robust second tier, and (c) TEACHSUP, COGACRCO, ANXMAT, and PERSEVAGR exert comparatively limited influence. Importantly, the results document variable contributions at the levels of both magnitude and rank without assuming cross-model comparability of absolute scales. The stable prominence of home resources and mathematics self-efficacy provides the empirical basis for the subsequent discussion of analytically salient predictors, while not implying that these variables constitute directly validated intervention targets.

DISCUSSION

This study set out to examine whether a pretrained tabular foundation model, TabPFN, can deliver competitive classification performance on PISA 2022 mathematical literacy data while producing interpretable predictor-importance patterns in a large-scale educational assessment

context (Hollmann *et al*., 2023; Hollmann *et al*., 2025). Regarding RQ1 and RQ2, the findings indicate that TabPFN achieved the most favorable overall balance of ROC–AUC, accuracy, precision, recall, and F1 on the PISA 2022 Türkiye mathematics sample (Table 1), with statistically significant gains over LR, LightGBM, and RF. At the same time, its performance was essentially indistinguishable from stacking and differed only marginally and non-significantly from XGBoost (Table 2). This pattern is consistent with broader evidence that several high-capacity learners tend to converge in performance on tabular prediction tasks, suggesting model pluralism rather than a single dominant algorithm in educational applications (Breiman, 2001; Chen & Guestrin, 2016; Probst *et al*., 2019; Romero & Ventura, 2020). In deployments where continuity with existing software ecosystems and accumulated expertise is important, well-tuned tree-based ensembles therefore remain a defensible choice (Biau & Scornet, 2016; Breiman, 2001). In settings where maintenance costs, periodic retraining, and limited tuning capacity are key constraints, TabPFN's no-tuning regime can offer a favorable trade-off between predictive performance and implementation effort for educational computing (Hollmann *et al*., 2023; Hospedales *et al*., 2022; Vanschoren, 2019).

These performance claims should therefore be read within the study's classification-oriented and threshold-based design. The results support the conclusion that TabPFN is a competitive model for distinguishing students around the PISA Level 2 proficiency benchmark; they do not, by themselves, establish that TabPFN is uniformly superior across all possible educational prediction tasks or that it is ready for direct operational deployment without further validation.

Turning to RQ3, the model-agnostic importance analyses revealed a stable and theory-consistent structure across models. HOMEPOS and MATHEFF consistently occupied the top positions, followed by CURIOAGR and DISCLIM. Parental education and family and peer *support* contributed at intermediate levels, whereas teacher support, cognitively demanding classroom activities, mathematics anxiety, and perseverance exerted comparatively modest influence (Tables 3 – 4; Appendix A, Figure S3; Akyüz & Pala, 2010; Kamaliyah *et al*., 2013; OECD, 2023). Because permutation importance and SHAP values rely on different definitions and scales, these results are interpreted primarily in terms of rank patterns rather than absolute magnitudes (Lundberg & Lee, 2017; Molnar, 2025). Consistent with RQ3, the interpretation of these findings is limited to identifying the variables that showed the strongest and most stable predictive relevance across the evaluated models. Importantly, importance scores capture associations rather than causal effects. Therefore, the present findings should not be read as evidence that changing any single predictor would directly improve mathematical literacy; rather, they indicate which contextual and learner-related variables were most consistently informative for the classification task.

The predictor pattern can be linked more critically to established perspectives in PISA-based mathematics achievement research. HOMEPOS reflects the material and cultural resources available for learning, whereas MATHEFF is consistent with prior evidence that mathematics-related efficacy beliefs are closely associated with achievement. CURIOAGR and DISCLIM further connect the model results to motivational engagement and classroom climate. At the same time, the lower ranking of teacher support, cognitive activation, anxiety, and perseverance should not be interpreted as evidence that these constructs are theoretically unimportant. Their lower importance in this study may reflect the binary thresholded outcome, overlap among questionnaire indices, self-report measurement limitations, or the restricted student-level modelling framework used here (Bakan Kalaycıoğlu, 2015; Mislevy, 2018; OECD, 2023; Rutkowski *et al*., 2010; Wang *et al*., 2023).

Implications for Educational Computing and Decision Support

For educational computing research, the findings suggest that TabPFN-style foundation models can be evaluated not only by their classification performance but also by the stability and interpretability of their predictor-importance profiles. This implication remains methodological

rather than prescriptive: the study does not test an intervention model, but it shows that interpretable machine learning outputs can help researchers examine whether model predictions are driven by educationally meaningful constructs. The joint prominence of HOMEPOS and MATHEFF indicates that material learning resources and mathematics-related efficacy beliefs carried the strongest predictive signal in the present classification task. This pattern is consistent with prior work emphasizing the role of socioeconomic resources and motivational factors in mathematics achievement (Akyüz & Pala, 2010; OECD, 2023). Accordingly, these variables may be viewed as analytically salient predictors within the model, rather than as directly validated intervention targets.

Explainability remains central to the interpretation of model outputs in educational assessment. For teachers, researchers, and school leaders to evaluate model-based results critically, explanations need to be concise, transparent, and connected to familiar educational constructs. The use of out-of-fold SHAP summaries for TabPFN and permutation importance for the baseline models made it possible to compare whether different algorithms relied on similar predictor patterns. In this study, the repeated prominence of HOMEPOS and MATHEFF, followed by CURIOAGR and DISCLIM, suggests that the models did not rely on arbitrary or purely technical signals; instead, the highest-ranking variables were substantively interpretable within the broader PISA framework (Lundberg & Lee, 2017; Molnar, 2025; OECD, 2023). This finding strengthens the methodological value of the study by showing that competitive performance was accompanied by a coherent predictor-importance structure.

Thresholding and decision rules should also be interpreted cautiously. Aggregate performance metrics can mask asymmetries between error types and between student subgroups. Although the present study used a binary outcome based on a meaningful proficiency threshold, its primary contribution lies in comparing model performance and predictor importance rather than establishing an operational early-warning system. Future applications that use predicted probabilities for resource allocation or individual-level decisions would require additional analyses, including calibration checks, cost-sensitive thresholds, and subgroup-level validation (Chicco & Jurman, 2020; Saito & Rehmsmeier, 2015; Sokolova & Lapalme, 2009). Thus, the present results should be regarded as a methodological basis for further validation, not as a complete decision-support framework.

From an implementation perspective, TabPFN's single-step inference reduces the overhead associated with hyperparameter tuning and can be aligned with periodic model evaluation cycles in educational data analysis (Hollmann *et al*., 2023; Hollmann *et al*., 2025). By contrast, stacking and boosting pipelines often require more frequent retuning but benefit from mature software ecosystems and well-understood diagnostics for feature effects (Chen & Guestrin, 2016; Probst *et al*., 2019). The practical relevance of TabPFN in this study therefore lies in its performance-to-effort profile and its capacity to yield interpretable predictor rankings under a consistent evaluation pipeline. This position supports a model-pluralist interpretation: TabPFN is not presented as a replacement for established learners, but as a competitive alternative that may be useful when analytic simplicity and reproducibility are important concerns.

Accordingly, any practical implication drawn from these findings should be understood as a basis for further validation rather than as evidence of immediate decision-support effectiveness. The study did not test dashboards, intervention assignments, or resource-allocation rules; therefore, claims about educational decision-making are limited to the methodological potential of combining benchmarked classification performance with interpretable predictor profiles (Chicco & Jurman, 2020; Sokolova & Lapalme, 2009).

Positioning the Contribution and Future Directions

Conceptually, the study links methodological performance with interpretability in large-scale educational assessment. In relation to RQ1 and RQ2, the results demonstrate that a pretrained tabular model can achieve performance on PISA 2022 mathematical literacy data that is

comparable to state-of-the-art tree-based ensembles without task-specific hyperparameter tuning (Hollmann *et al*., 2023; Hollmann *et al*., 2025; Romero & Ventura, 2020). In relation to RQ3, the study further shows that the most influential variables were not model-specific anomalies, but relatively stable predictors across multiple learners. The recurring importance of home resources, mathematics self-efficacy, curiosity, and disciplinary climate indicates that the classification models drew on constructs that are substantively meaningful within PISA-based mathematics achievement research (Holmes *et al*., 2019; OECD, 2023). This contributes to the literature by combining model benchmarking with an explicit examination of predictor-importance consistency, thereby moving beyond accuracy-only comparisons.

The contribution to scientific knowledge is therefore methodological and interpretive. First, the study extends PISA-based educational data-mining research by evaluating a tabular foundation model in a major-domain large-scale assessment context rather than relying only on conventional classifiers. Second, it shows that TabPFN's performance should be interpreted through a model-pluralist lens because its advantage over the strongest baselines was small and not statistically reliable in every comparison. Third, it demonstrates that interpretability evidence can be examined for cross-model consistency, allowing predictor-importance results to be evaluated as stable, theory-aligned classification signals rather than as isolated outputs of a single algorithm. In this respect, the study moves the literature beyond accuracy-only benchmarking toward assessment-informed, interpretable model evaluation.

At the same time, the relatively low importance estimates for some teacher- and classroom-related constructs should be interpreted cautiously. Limited measurement granularity, reliance on single-occasion self-reports, and the cross-sectional nature of PISA can attenuate their apparent contributions (Mislevy, 2018; Rutkowski *et al*., 2010). Accordingly, low feature importance should not be interpreted as evidence that these constructs are unimportant for learning in general. Rather, it indicates that, within the present dataset and modelling framework, they contributed less to the classification of mathematical literacy than home resources and mathematics self-efficacy. Future work should therefore complement cross-validated estimates from large-scale assessments with richer classroom instrumentation, longitudinal traces that capture change over time, and group-aware cross-validation schemes that respect school-level clustering (OECD, 2009; Romero & Ventura, 2020). Extending the present approach to fairness-sensitive evaluations would further strengthen the responsible use of machine learning models in educational assessment contexts.

Limitations and Future Research

Although TabPFN demonstrated high discriminative ability and balanced accuracy for classifying mathematics achievement in the PISA 2022 Türkiye data, several limitations related to scope and design qualify the conclusions and indicate directions for further work.

Scope, sampling, and generalizability. The analyses focused exclusively on the Türkiye subsample of PISA 2022. Cross-national differences in curricula, assessment systems, and student populations may alter model behavior and thus constrain the external validity of the findings (OECD, 2009; Rutkowski *et al*., 2010). In addition, because the primary objective was algorithmic benchmarking rather than population parameter estimation, PISA sampling weights were not incorporated into model training or evaluation. Future research should therefore examine the sensitivity of TabPFN and ensemble-model performance to weight-adjusted modelling pipelines. A natural extension is to evaluate out-of-sample and out-of-domain performance across multiple countries and PISA cycles (e.g., 2018, 2025) and to implement cluster-aware validation schemes—such as school-stratified group K-fold or leave-one-school-out procedures—to reduce leakage associated with clustered sampling.

Outcome definition and dichotomisation. In this study, mathematics achievement was represented as a binary label derived from a continuous PISA proficiency score. This decision was consistent with the study's classification-oriented aim and with the use of proficiency

thresholds in large-scale assessment reporting. Nevertheless, dichotomising a continuous proficiency scale inevitably reduces score variability and may obscure differences among students located near or far from the cut-off. Therefore, the findings should be interpreted as evidence about classification performance around a meaningful proficiency threshold rather than as a complete model of continuous mathematics achievement. Future research should extend RQ1–RQ2 to continuous or ordinal outcomes, such as predicted proficiency scores or PISA proficiency levels, and compare whether the relative performance of TabPFN and the baseline models remains stable across classification, regression, and ordinal modelling frameworks.

Measurement and construct validity. Several predictors are derived from single-occasion self-reports and may be affected by limited granularity, response styles, or context-specific shifts, all of which can attenuate observed associations (Mislevy, 2018; Holmes *et al*., 2019). Replications that incorporate richer classroom instrumentation, multiple informants, and longitudinal traces would better align importance profiles with learning theories and help mitigate measurement error (OECD, 2023).

Missing data and preprocessing. Because TabPFN requires fully observed numeric input, the current pipeline relied on within-fold imputation to avoid information leakage. Nonetheless, the results may be sensitive to this particular imputation strategy. Future work should systematically compare alternative approaches—such as multiple imputation, kNN-based methods, and missingness-aware encodings—within the cross-validation loop, and report robustness checks and uncertainty estimates around performance and importance scores (Molnar, 2025; Witten *et al*., 2016; Zhang, 2012).

Evaluation design and inference. Model comparisons were based on stratified 10-fold cross-validation, and fold-level AUCs were analyzed using a Friedman test with Holm-adjusted pairwise contrasts. Although this resampling-aware design reduces optimistic bias, dependence across folds may still affect inferential precision. Repeated cross-validation, external hold-out evaluations (e.g., temporal or cross-cycle), or corrections explicitly designed for resampling-based inference (such as the Nadeau–Bengio adjustment) could further strengthen the robustness of conclusions (Nadeau & Bengio, 2003; Probst *et al*., 2019).

Fairness, shift, and subgroup validity. The present study did not conduct a formal fairness audit. Future research should examine group-wise error rates (e.g., equalized odds), monitor distributional shift across cohorts or years, and investigate domain adaptation strategies for cross-cycle generalization (Kizilcec & Lee, 2022; OECD, 2023). Such analyses would align TabPFN-style models more closely with equity and responsible-AI agendas in education.

Compute, sustainability, and deployment. Although TabPFN's strong performance without task-specific tuning is operationally attractive, the study did not report detailed metrics on computation time, latency, or energy use relative to tree-based baselines. Systematic reporting and benchmarking along these dimensions would provide a clearer basis for large-scale adoption. Comparative audits that jointly consider accuracy, sustainability, and transparency—for instance, including explainability artifacts for teacher-facing dashboards—would better inform institutional choices (Lundberg & Lee, 2017; Molnar, 2025).

In sum. While the current evidence highlights TabPFN's advantages over strong baselines and its near-parity with top-performing competitors, establishing its broader role in educational computing will require cross-country and cross-cycle replications, missing-data-aware modeling pipelines, calibration and cost-sensitive metrics, validation schemes that respect clustering, and fairness-aware evaluations.

These extensions would help determine whether the predictor-importance patterns observed in this study remain stable across different mathematical literacy assessment contexts and modelling conditions.

CONCLUSION

This study provides converging evidence that a pretrained tabular model, TabPFN, can achieve high discriminative performance and balanced classification of students' mathematics achievement in the PISA 2022 Türkiye sample. Using stratified 10-fold cross-validation, TabPFN obtained the strongest scores on the primary metrics, and Holm-adjusted pairwise comparisons of fold-level ROC–AUC confirmed statistically significant advantages over LR, LightGBM, and RF, performance parity with stacking, and a small, non-significant gap relative to XGBoost. Taken together, the findings for RQ1 and RQ2 position TabPFN as a competitive alternative to widely used baselines in educational data mining (Breiman, 2001; Chen & Guestrin, 2016; Hollmann *et al.*, 2023; Hollmann *et al.*, 2025; Probst *et al.*, 2019; Romero & Ventura, 2020).

These conclusions are intentionally bounded by the evidence generated in this study. The findings support competitive classification and stable predictor-importance patterns for the PISA 2022 Türkiye mathematical literacy sample, but they do not support causal claims about the predictors or direct recommendations for individual-level interventions.

With respect to RQ3, model-agnostic importance analyses yielded a coherent and theory-consistent profile. Home resources and socioeconomic background and mathematics self-efficacy systematically occupied the top ranks; curiosity and interest in new knowledge and classroom disciplinary climate formed a recurrent second tier; parental education and parental and peer support showed intermediate contributions; and teacher support, cognitive activation, mathematics anxiety, and perseverance contributed only modestly. These findings directly address the third research question by identifying the variables that were most influential in predicting mathematical literacy across the evaluated models. The consistency of the rankings suggests that strong predictive performance was accompanied by an interpretable and educationally meaningful predictor profile, while the associational nature of feature importance requires caution against causal or intervention-level interpretations (Lundberg & Lee, 2017; Molnar, 2025; OECD, 2023).

From an operational standpoint, TabPFN attains these results without task-specific hyperparameter tuning, yielding a favourable performance-to-effort profile for large-scale educational analytics. At the same time, its close proximity to stacking and XGBoost reinforces a model-pluralist view: in settings where transparency and continuity with existing infrastructure are paramount, tree-based ensembles remain a defensible option, whereas in scenarios where periodic deployment and low maintenance are key constraints, TabPFN represents a pragmatic choice (Breiman, 2001; Chen & Guestrin, 2016; Hollmann *et al.*, 2023; Hollmann *et al.*, 2025; Romero & Ventura, 2020).

In summary, TabPFN emerges as a competitive approach for classifying mathematical literacy in the PISA 2022 Türkiye sample, combining strong discrimination with a stable and interpretable predictor-importance structure. This interpretation of RQ3 narrows the contribution to the identification of influential predictors rather than making direct claims about targeted support processes. Accordingly, the conclusions should be understood as specific to PISA 2022 mathematical literacy and should not be generalized to the entire PISA assessment without additional domain-specific evidence. Future research should consolidate these findings through cross-country and cross-cycle validation, incorporate calibration and cost-sensitive evaluation when models are intended for decision-making use, and examine robustness under clustered sampling and distributional shift (Chicco & Jurman, 2020; Nadeau & Bengio, 2003; OECD, 2009; Saito & Rehmsmeier, 2015).

Overall, the main scientific contribution of the study is to show that tabular foundation models can be evaluated in large-scale educational assessment through a joint performance-and-interpretability lens. This provides a more cautious and educationally meaningful basis for future research on machine learning in assessment than performance metrics alone.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Data Availability Statement

The data used in this study are publicly available from the OECD PISA 2022 Database. The dataset, codebooks, questionnaires, and related documentation can be accessed through the official OECD PISA 2022 data repository: https://www.oecd.org/en/data/datasets/pisa-2022-database.html

Informed Consent

Not applicable. This study used publicly available, anonymised secondary data from the PISA 2022 database and did not involve direct data collection from human participants by the authors.

Contribution of Authors

B.E.: Conceptualization, Literature Review, Methodology, Investigation, Visualization, Formal Analysis, Writing – original draft, Writing – review & editing. **T.N.:** Methodology, Data Curation, Software, Formal Analysis, Validation, Resources, Writing – review & editing.

Orcid

Büşra Eren https://orcid.org/0000-0001-7565-1025
Tuğba Niksarlı https://orcid.org/0009-0001-4445-2311

REFERENCES

Akyüz, G., & Pala, N.M. (2010). PISA 2003 sonuçlarına göre öğrenci ve sınıf özelliklerinin matematik okuryazarlığına ve problem çözme becerilerine etkisi [The effect of student and class characteristics on mathematics literacy and problem solving in PISA 2003]. *Elementary Education Online*, *9*(2), 668–678. https://ilkogretim-online.org/index.php/pub/article/view/761

Baker, R.S.J.d., & Inventado, P.S. (2014). Educational data mining and learning analytics. In J.A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 61–75). Springer. https://doi.org/10.1007/978-1-4614-3305-7_4

Bakan Kalaycıoğlu, D. (2015). The influence of socioeconomic status, self-efficacy, and anxiety on mathematics achievement in England, Greece, Hong Kong, the Netherlands, Turkey, and the USA. *Educational Sciences: Theory & Practice*, *15*(5), 1391-1401. https://doi.org/10.12738/estp.2015.5.2731

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cheema, J.R., & Kitsantas, A. (2014). Influences of disciplinary classroom climate on high school student self-efficacy and mathematics achievement: A look at gender and racial–ethnic differences. *International Journal of Science and Mathematics Education*, *12*(5), 1261-1279. https://doi.org/10.1007/s10763-013-9454-4

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), Article 6. https://doi.org/10.1186/s12864-019-6413-7

Hanushek, E.A., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E.A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Elsevier. https://doi.org/10.1016/B978-0-444-53429-3.00002-8

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., …, Oliphant, T.E. (2020). Array

programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hollmann, N., Müller, S., Eggensperger, K., & Hutter, F. (2023). TabPFN: A transformer that solves small tabular classification problems in a second. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview. https://doi.org/10.48550/arXiv.2207.01848

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S.B., Schirrmeister, R.T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, *637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign. https://curriculumredesign.org/our-work/artificial-intelligence-in-education/

Hosmer, D.W., Jr., Lemeshow, S., & Sturdivant, R.X. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5149–5169. https://doi.org/10.1109/TPAMI.2021.3079209

Kamaliyah, K., Zulkardi, Z., & Darmawijoyo, D. (2013). Developing the sixth level of PISA-like mathematics problems for secondary school students. *Journal on Mathematics Education*, *4*(1), 9–28. https://doi.org/10.22342/jme.4.1.562.9-28

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 3146-3154). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kizilcec, R.F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J.P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 2390–2395). Association for Computing Machinery. https://doi.org/10.1145/2858036.2858402

Kizilcec, R.F., & Lee, H. (2022). Algorithmic fairness in education. In W. Holmes & K. Porayska-Pomsta (Eds.), *The ethics of artificial intelligence in education* (Chap. 10). Routledge. https://doi.org/10.4324/9780429329067-10

Liu, X., Yang Hansen, K., Valcke, M., & De Neve, J. (2024). A decade of PISA: Student perceived instructional quality and mathematics achievement across European countries. *ZDM – Mathematics Education, 56*, 859–891. https://doi.org/10.1007/s11858-024-01630-7

Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*. https://doi.org/10.48550/arXiv.1705.07874

Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage.

Mislevy, R.J. (2018). *Sociocognitive foundations of educational measurement*. Routledge. https://doi.org/10.4324/9781315269473

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

OECD. (2009). PISA data analysis manual: SPSS (2nd ed.). OECD Publishing. https://doi.org/10.1787/9789264056275-en

OECD. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,

Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications, 41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Probst, P., Wright, M.N., & Boulesteix, A.L. (2019). Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge *Discovery*, *9*(3), e1301. https://doi.org/10.1002/widm.1301

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), Article e1355. https://doi.org/10.1002/widm.1355

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, *39*(2), 142–151. https://doi.org/10.3102/0013189X10363170

Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), Article e0118432. https://doi.org/10.1371/journal.pone.0118432

Seni, G., & Elder, J.F. (2010). *Ensemble methods in data mining: Improving accuracy through combining predictions.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00240ED1V01Y200912DMK002

Sill, J., Takács, G., Mackey, L., & Lin, D. (2009). *Feature-weighted linear stacking* [Preprint]. arXiv. https://arxiv.org/abs/0911.0460

Slater, S., Joksimovic, S., Kovanović, V., Baker, R.S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85–106. https://doi.org/10.3102/1076998616666808

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437. https://doi.org/10.1016/j.ipm.2009.03.002

Ting, K.M., & Witten, I.H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, *10*, 271–289. https://doi.org/10.1613/jair.594

van der Laan, M.J., Polley, E.C., & Hubbard, A.E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), Article 25. https://doi.org/10.2202/1544-6115.1309

Vanschoren, J. (2019). Meta-learning. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated machine learning: Methods, systems, challenges* (pp. 35–61). Springer. https://doi.org/10.1007/978-3-030-05318-5_2

Wang, X., Perry, L.B., & Malpique, A. (2023). Factors predicting mathematics achievement in PISA: A systematic review. *Large-scale Assessments in Education, 11*, Article 24. https://doi.org/10.1186/s40536-023-00174-8

Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann. https://doi.org/10.1016/C2015-0-02071-8

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

Zhang, S. (2012). Nearest neighbor selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), 2541–2552. https://doi.org/10.1016/j.jss.2012.05.073

APPENDICES

Appendix A. Supplementary Figures and Tables

The appendix provides additional diagnostic outputs that support the model-comparison and interpretation results reported in the main manuscript.

Supplementary Figure S1. Fold-Wise ROC–AUC Comparison Across Models

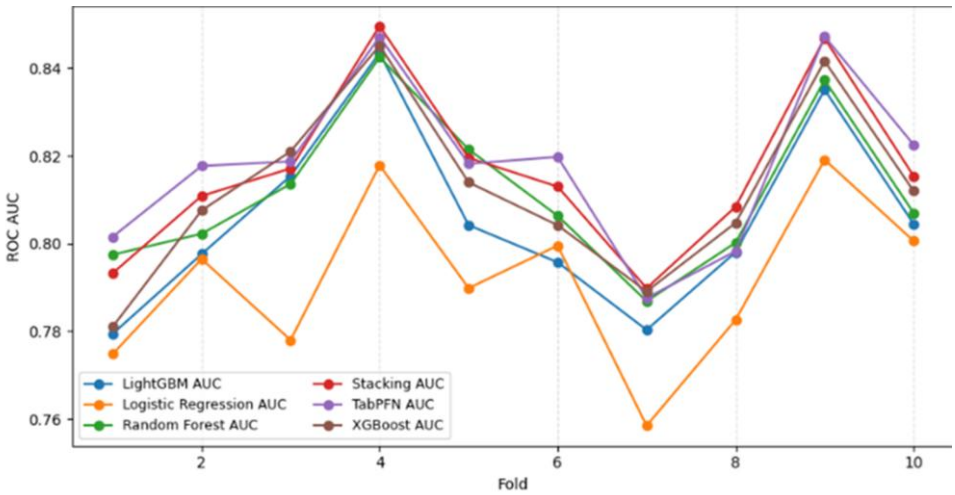

Figure S1. *Fold-wise ROC–AUC comparison across six models under stratified 10-fold cross-validation (PISA 2022 Türkiye sample).*

Note. Points denote fold-level ROC-AUC values; lines are used only to aid visualisation. Fold means (± *SD*) are reported in Table 1.

Supplementary Table S1. Confusion Matrices

Table S1. *Confusion matrices for the six models at the default 0.50 decision threshold.*

Model	Actual: Negative / Predicted: Negative	Actual: Negative / Predicted: Positive	Actual: Positive / Predicted: Negative	Actual: Positive / Predicted: Positive
LR	2823	995	1082	2350
RF	2956	862	1053	2379
XGBoost	2837	981	1032	2400
LightGBM	2948	870	1021	2411
Stacking	2818	1000	1093	2339
TabPFN	2952	866	995	2437

Note. All matrices were computed using the fixed 0.50 decision threshold used for the main model-comparison results.

Supplementary Figure S2. Model-Specific ROC Curves

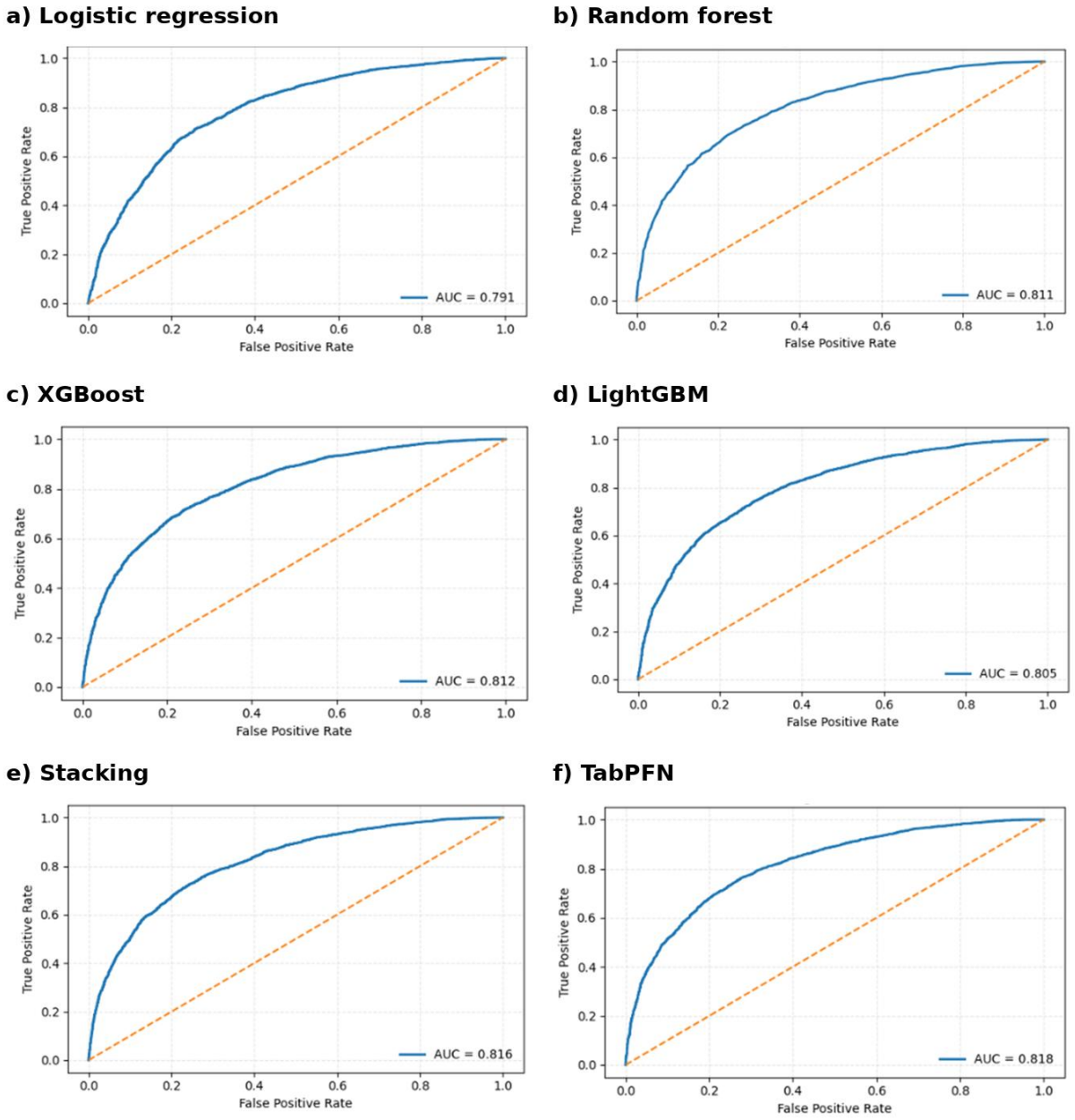

Figure S2. *Receiver operating characteristic (ROC) curves for individual models under stratified 10-fold cross-validation.*

Note. ROC-AUC is displayed in each panel; the dashed line indicates the no-discrimination baseline.

Supplementary Figure S3. Feature Importance by Model

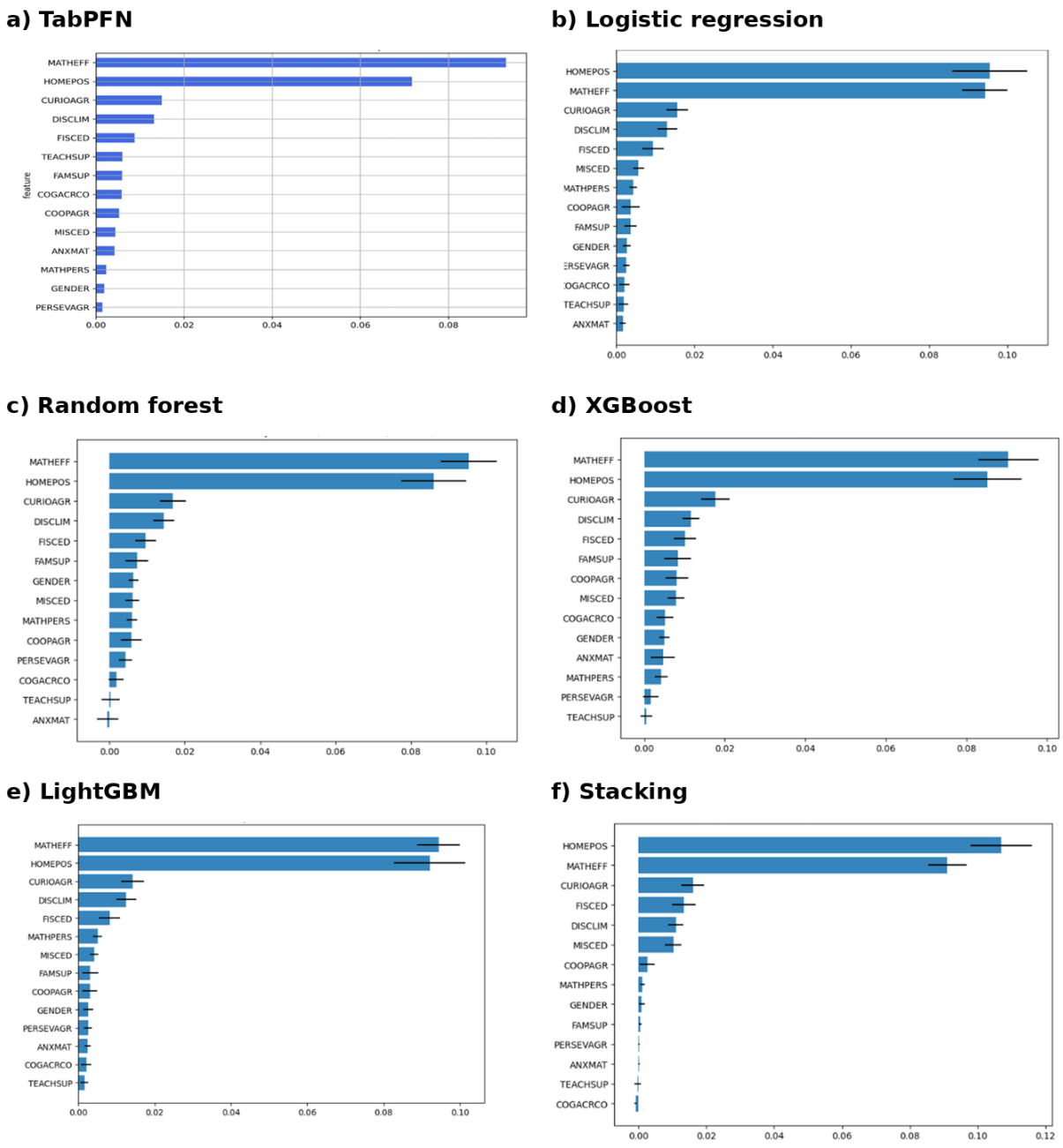

Figure S3. *Feature-importance profiles for the evaluated models.*

Note. For logistic regression, random forest, XGBoost, LightGBM, and stacking, bars indicate the mean decrease in ROC–AUC when the corresponding feature is randomly permuted. For TabPFN, bars represent mean absolute SHAP values computed on out-of-fold predictions. Error bars summarize variability across folds.