JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

When Similar Models Disagree: Model Multiplicity and Geographic Robustness in School Meal Utilization Prediction

Amanatullah Pandu Zenklinov [1,*], Andriano [1], Wisnowan Hendy Saputra [2], Marcella Ariani [1], Nabilla Rida Tri Nisa [3], and Sabrina Julietta Arisanty [4]

[1] Statistics Department, Universitas Matana, Tangerang 15810, Banten, Indonesia; email: amanatullah.zenklinov@matanauniversity.ac.id; andriano@matanauniversity.ac.id; marcella.ariani@student.matanauniversity.ac.id
[2] Computer Science Department, Bina Nusantara University, Jakarta 11530, Indonesia; email: wisnowan.saputra@binus.ac.id
[3] Multimedia Engineering Technology Study Program, Politeknik Negeri Media Kreatif Jakarta, Jakarta 12640, Indonesia; email: nabilla.nisa@polimedia.ac.id
[4] Department of Biostatistics, Epidemiology, and Public Health, Faculty of Medicine, Public Health, and Nursing, Universitas Gadjah Mada, Sleman 55281, Yogyakarta, Indonesia; email: sabjulietta@gmail.com
* Corresponding Author : Amanatullah Pandu Zenklinov

Abstract: Competing predictive models can achieve similar aggregate performance while producing different classifications and downstream conclusions for the same individuals. This model disagreement raises an important question about whether conventional performance metrics adequately capture the reliability of model-based analyses. This study investigates model-choice disagreement across multiple analytical layers using voluntary-response survey data from 8,802 respondents who reported receiving meals through Indonesia's Free Nutritious Meals program. Logistic Regression (LR), Random Forest, and XGBoost were evaluated using a nested school-disjoint validation framework. All three models were compared in terms of discrimination and probability calibration, while the competitive models were further examined through high-risk classification, predictive attribution, model-based scenario analysis, and province-holdout robustness. LR and XGBoost achieved nearly identical PR-AUC values (0.7270 and 0.7277, respectively), with similarly close pooled PR-AUC under province holdout (0.7288 and 0.7285). Despite this aggregate similarity, the models assigned different high-risk classifications to 5.41% of respondents. Their scenario analyses also produced different threshold-crossing rates: 8.87% of eligible respondents under LR (382 of 4,307) and 6.66% under XGBoost. Random Forest exhibited weaker probability calibration in the fitted configuration and was not retained for downstream comparison. Overall, the findings reveal that model agreement at the population level does not necessarily extend to individual classifications and scenario outputs. LR was selected as the reference model through documented judgment considering predictive performance, calibration, robustness, parsimony, and interpretability. These results demonstrate the importance of evaluating model-choice robustness beyond conventional predictive performance metrics, particularly when model outputs inform downstream interpretation or classification.

Keywords: Geographic Robustness; Grouped Validation; Model Multiplicity; Model-Choice Robustness; Probability Calibration; Scenario Analysis; School Feeding; School Meal Utilization.

Received: September, 21st 2026
Revised: October, 28th 2026
Accepted: October, 29th 2026
Published: November, 10th 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) licenses (https://creativecommons.org/licenses/by/4.0/)

1. Introduction

Predictive models with similar out-of-sample performance do not necessarily produce the same downstream conclusions [1], [2]. Two models can achieve nearly identical discrimination while assigning different probabilities to individual observations, producing different classifications at operational thresholds, emphasizing different predictive features, or responding differently to changes in model inputs. This distinction is important when predictive outputs are used beyond performance benchmarking, because similar aggregate performance does not establish that competing models are interchangeable. The issue is particularly

relevant to structured tabular data, where models with different structures can achieve comparable discrimination despite producing different predictions and analytical outputs [3], [4]. In this study, model multiplicity refers to the coexistence of similarly performing models that disagree on individual predictions or downstream outputs, while model-choice robustness refers to the extent to which analytical findings remain consistent when one competitive model is replaced by another.

Previous studies have documented disagreement among predictive models at several analytical levels. Despite similar population-level accuracy, different models have assigned different individuals to high-risk groups. For example, three breast cancer risk models identified different women as belonging to the high-risk group [5], while logistic regression and two XGBoost variants with comparable discrimination in a cardiovascular cohort showed only partial overlap among individuals ranked at the highest risk [6]. Research on predictive multiplicity and underspecification has further demonstrated that multiple well-performing models can produce different predictions for the same observations [1], [2]. Such disagreement is not limited to predictive outputs: studies in explainable machine learning have reported inconsistencies among explanatory outputs [6]. while research on counterfactual explanations has shown that similarly accurate models can generate different recommended changes [7]. These findings indicate that aggregate predictive performance alone cannot establish whether competing models produce consistent individual classifications, emphasize the same predictive features, or respond similarly to changes in input variables.

Although previous studies have established that disagreement can occur across different stages of predictive modeling, it remains unclear whether similarly performing models produce consistent findings across multiple downstream analytical layers when evaluated within a common framework. In particular, comparable discrimination does not necessarily imply agreement in individual-level classification, predictive attribution, or model-based scenario analysis. Assessing such agreement also requires attention to the evaluation setting. When observations are clustered, record-wise validation may overestimate performance on previously unseen groups [8]. Moreover, probability-based decisions require consideration of calibration alongside discrimination [9], while geographic generalization can be examined by systematically withholding natural geographic units during validation [10], [11]. These considerations suggest that model-choice robustness should not be inferred from aggregate predictive performance alone. The central concern is whether apparent equivalence in aggregate performance masks consequential disagreement in individual-level decisions and subsequent analytical findings. Such agreement cannot be assumed from comparable discrimination; it must be examined empirically across the different outputs on which model-based conclusions may depend. Accordingly, model-choice robustness requires examining both downstream disagreement and geographic robustness to determine whether conclusions remain consistent across competitive models and different validation settings.

The school meal utilization setting in Indonesia provides an empirical context for examining these issues. The Free Nutritious Meals program (*Makan Bergizi Gratis*, MBG), whose target groups include students [12], is the program from which the survey respondents report receiving meals. The respondent-reported data provide an opportunity to examine model multiplicity in predicting persistent low meal utilization, an outcome defined at the individual respondent level. The data also have a clustered structure across schools and geographic regions, making them relevant for evaluating model agreement under school-disjoint validation and geographic holdout. Accordingly, MBG serves as the empirical application through which model multiplicity, downstream disagreement, and geographic robustness are investigated. The study does not evaluate the effectiveness of the MBG program or the performance of individual schools or *Satuan Pelayanan Pemenuhan Gizi* (SPPG) units.

Against this background, this study investigates when and where similarly performing predictive models disagree. Specifically, it examines whether models with comparable predictive performance produce consistent results in individual-level classification, predictive attribution, and model-based scenario analysis, and whether their predictive performance remains comparable under geographic holdout evaluation. The main contributions of this study are as follows:

- It examines model-choice robustness across multiple analytical layers, including individual classifications, predictive attribution, and model-based scenario outputs.
- It provides empirical evidence that models with similar predictive performance can produce different individual classifications and scenario outcomes.

- It evaluates the geographic robustness of competing models through province-holdout validation, complementing the school-disjoint evaluation.

Together, these contributions distinguish similarity in aggregate predictive performance from agreement in downstream model outputs and geographic generalization. The analysis therefore moves beyond identifying the best-performing model to examining the extent to which conclusions derived from competitive models can be considered consistent or model-dependent.

2. Literature Review

2.1. School-Meal Utilization and Predictive Analytics

Prior research on school-meal use has focused mainly on two areas. One line of research uses regression models to examine participation: surveys have related participation to family, school, and perception variables among German schoolchildren [13], US middle-school students [14], Pacific Islander parents [15] and, in multilevel models, US students [16]. Their outcomes differ from the persistent low utilization examined here. Related outcomes have also been reviewed: acceptability [17] and plate food waste [18].

A second line of work applies machine learning to related nutrition and school-feeding outcomes: comparisons of random forest, support vector machine, linear discriminant analysis and logistic regression for undernutrition among Philippine schoolchildren [19], and classification trees for municipal compliance with a family-farming purchase requirement in Brazil's school feeding program [20]. Recent work on Indonesia's MBG program includes a commentary on food-safety risks [21] and machine-learning studies of menu-level nutritional adequacy [22] and nutrient-consumption level [23]. Closest to the outcome studied here, a study of 132 students modeled daily MBG consumption from satisfaction and perception items and compared Random Forest and XGBoost [24].

Together, these studies demonstrate the feasibility of modeling school-meal participation and related consumption outcomes, but they leave open whether the resulting conclusions remain stable across competing models, validation protocols, and geographic settings.

2.2. Comparative Machine-Learning Modeling

Comparative prediction studies often evaluate several model families, although more flexible models do not consistently outperform simpler alternatives. A systematic review of 71 clinical comparisons found no difference in discrimination (area under the receiver operating characteristic curve, AUC) between logistic regression and machine learning at low risk of bias (difference in logit(AUC) 0.00, 95% confidence interval [CI], −0.18 to 0.18), whereas comparisons at high risk of bias favored machine learning [3]. In a cardiovascular cohort, logistic regression had the smallest variability across data splits and the lowest Brier score against two XGBoost variants [6]. These results depend on the outcome, the predictors, and the evaluation protocol, and they do not establish a general ranking of model families.

The three families used here differ in structure. Logistic regression is a linear, parametric probabilistic model; random forests combine many trees grown on resampled data with random feature selection [25]; and XGBoost implements regularized gradient tree boosting [26]. For structured data with meaningful features, there is often no significant performance difference between complex classifiers and simpler ones [4]. Evaluating these contrasting structures under a common protocol allows conclusions that are robust to model family to be distinguished from those that are model-dependent.

2.3. Robust Evaluation Beyond Predictive Performance

Headline discrimination alone provides an incomplete basis for evaluating predictive models. Validation strategies must reflect the underlying data structure and intended application. When observations belong to clusters and predictions are intended for previously unseen clusters, record-wise cross-validation can overestimate predictive performance compared with subject-wise validation, as demonstrated in sensor-based clinical prediction [8]. Similarly, validation that ignores spatial structure can produce over-optimistic estimates in ecological mapping [27], [28]. Optimistic estimates may also arise when model configurations are selected using evaluation data; repeating configuration selection within each outer training set

helps prevent information from the evaluation data from influencing model selection [29]. Furthermore, when predicted probabilities are used for threshold-based classification or downstream decisions, calibration is essential because discrimination alone does not establish the reliability of predicted risks [9]. Nevertheless, calibration was not addressed in 56 of 71 clinical comparisons[3].

Similar predictive performance does not necessarily imply similar downstream behavior. Predictive multiplicity refers to the ability of a prediction problem to admit competing models that produce conflicting predictions [1], while an underspecified pipeline may yield multiple predictors with similarly strong held-out performance but different downstream behavior [2]. For example, among 31,115 women, the BCRAT, BCSC, and IBIS breast cancer risk models showed similar population-level accuracy, yet more than 21% of women were classified as high risk by one model and average risk by another at a 1.67% threshold [5]. Similarly, in a cardiovascular cohort, logistic regression and two XGBoost variants achieved comparable discrimination, but risk-ranking correlations with logistic regression of approximately 0.95 and 0.83 corresponded to overlaps of only about 71% and 50%, respectively, in the highest-risk decile [6]. These examples illustrate that agreement in aggregate performance does not guarantee agreement in risk rankings or classifications, let alone across the multiple stages of a downstream analytical pipeline involving thresholds, classifications, predictive attributions, and scenario analyses. Performance may also vary across geographic or institutional settings. Internal-external validation addresses this issue by iteratively withholding natural units, such as sites or regions, to assess generalization across settings, although it does not replace independent external validation [11]. This approach has also been used to compare competing models in clustered datasets [10]. For respondent-level prediction, an important unresolved question is therefore whether analytical conclusions remain stable when model choice, decision thresholds, and geographic settings vary.

2.4. Interpretability, Scenario Analysis, and Research Gap

Attribution methods such as SHapley Additive exPlanations (SHAP) assign a prediction to input features [30], with efficient forms for tree ensembles [31] and a closed form for linear models under feature independence [30]. Attributions describe a fitted model, not the data-generating process: Shapley-based importance has known limits [32], the choice of conditional distribution makes feature relevance a causal problem [33], prediction and causal explanation are distinct goals [34], and explanation methods often disagree.

Scenario analysis of a fitted model is related to counterfactual explanations and algorithmic recourse. A counterfactual explanation describes the smallest change to the inputs that alters a model's output [35]; recourse formulations restrict changes to actionable variables [36], and feasibility has been approached through data density and connection to the training data [37], [38]. Surveys distinguish contrastive explanations from consequential recommendations, which require causal knowledge that a predictive model does not supply [39], [40]. Counterfactual recommendations can differ across similarly accurate models [7], and applied work has coupled SHAP with constrained counterfactual scenarios for student dropout [41]. Because prediction and causal policy problems are distinct [42], a model-based scenario is a statement about a fitted model, not about the effect of a change in practice.

These distinctions also clarify how future work can move beyond predictive robustness. Cross-model explainable AI (XAI) comparisons can indicate whether leading predictive patterns remain stable across model classes, whereas establishing whether a potentially modifiable factor actually changes meal consumption requires a causal design or causal-inference framework [40], [42]. Predictive attribution and causal effect estimation should therefore be treated as complementary but distinct stages of future investigation [33], [34].

Conclusions that hold in aggregate need not hold at the other layers of a downstream analysis. Model comparison, calibration, grouped or geographic validation, explanation and counterfactual scenarios have been evaluated in other applications, often as separate components and in some cases within one design [41], [43], and agreement across similarly performing models has been examined for risk classifications, rankings and counterfactual recommendations [5]–[7]. Existing work therefore leaves unresolved whether the same competitive models remain stable across discrimination, calibration, individual classification, predictive attribution, scenario output, and geographic holdout within one evaluation framework. This motivates evaluating model-choice robustness across multiple layers rather than through a single performance metric.

3. Proposed Method

3.1. Overall Computational Framework

This study adopts a comparative machine-learning framework to predict persistent low meal utilization from respondent-reported survey data and examine the robustness of the resulting findings across different models and analytical settings. The analysis is predictive rather than causal, focusing on model performance and the consistency of model-derived outputs rather than estimating causal effects [34].

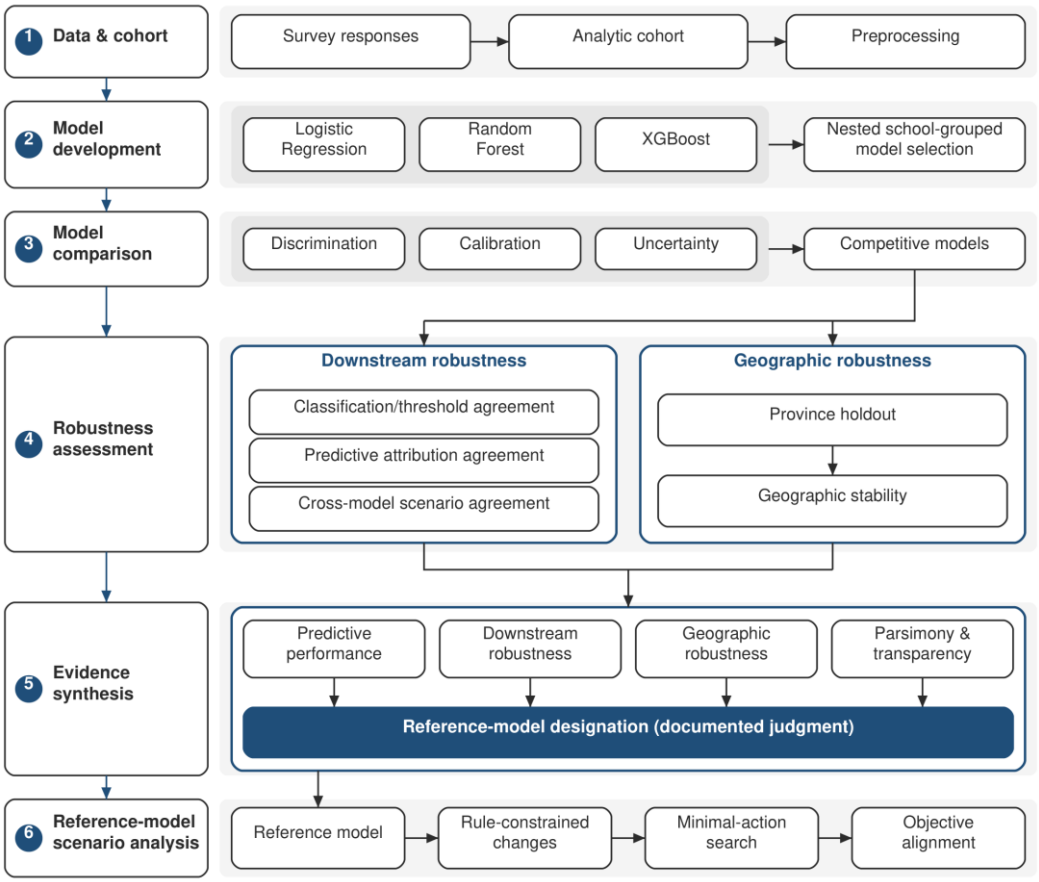


Figure 1. Overview of the six-stage study framework

As illustrated in Figure 1, the framework consists of six stages. First, the survey responses are processed to construct the analytic cohort and prepare the features for modeling. Second, Logistic Regression, Random Forest, and XGBoost are developed, with model configurations selected through nested school-grouped validation. Third, their predictive performance is compared using out-of-fold discrimination, calibration, and uncertainty estimates. Models demonstrating sufficient predictive competitiveness are then retained for two complementary robustness assessments. The first examines whether model-derived classifications, thresholds, feature attributions, and scenario results remain consistent across models. The second evaluates geographic robustness through province-holdout validation. Evidence from these assessments, together with considerations of model parsimony and transparency, is subsequently considered when designating a reference model. This designation reflects a documented analytical judgment rather than a formal selection rule or a claim of statistical superiority; no equivalence margin was prespecified. Finally, the selected reference model is used for the operational threshold analysis and scenario assessment involving rule-constrained changes to the input variables.

Two prediction contexts are distinguished throughout the analysis. Out-of-fold predictions are used to evaluate predictive performance, calibration, threshold-based classifications, model agreement, and feature attributions. These predictions are generated for respondents whose schools were excluded from the corresponding model-fitting process and are referred

to as held-out predictions in this study. In contrast, scenario analyses use predictions from a model fitted to the complete analytic dataset. Consequently, scenario baselines and resulting probabilities are in-sample model outputs and should not be interpreted as independent validation results.

3.2. Study Data and Operational Endpoint

The study used respondent-reported information collected through a structured questionnaire completed by students, their parents or guardians, or both together (item A01). Responses included in the analysis were recorded between 1 April and 15 September 2026. The respondent was the unit of analysis, while schools and provinces served as the grouping and validation units.

The questionnaire was administered online through Google Forms and distributed via school and community communication networks to students and/or their parents or guardians. The study employed a non-probability voluntary-response sampling approach, and participation was voluntary. The survey was not restricted to MBG recipients. Consequently, the sample had no defined sampling frame, selection probabilities, or sampling weights, and the data were not weighted to any target population. The questionnaire recorded the province, district, subdistrict, and school name, a respondent code (item A02), and a response timestamp. Of the 11,955 submissions received, 8,964 respondents reported currently receiving MBG meals (item A14 = "Ya"). The final analytic cohort consisted of 8,802 respondents who reported currently receiving MBG meals, had an observed endpoint, and had a valid canonical school identifier. This identifier was constructed deterministically from normalized province, district, and school-name fields, without fuzzy matching, and was used only for grouping purposes.

No formal ethics review or exemption determination was obtained before data collection. Based on the study records and institutional documentation available to the authors, no documented institutional determination established whether formal ethics review was required for this survey. Therefore, the study is not described as ethically approved, exempt, or formally determined not to require review. The questionnaire began with an information screen, followed by the question: "After reading the information above, are you willing to participate in this research?" All 11,955 submissions recorded electronic agreement to participate. No separate institutional permission, parental consent, or assent procedure was documented. The questionnaire did not collect respondents' names or contact details. Geographic and school information used for grouping was retained in the respondent-level dataset, with access restricted to the research team. The respondent-level data are not publicly released, and all results are reported in aggregate form.

Persistent low meal utilization was defined using two respondent-reported items. AT04 asked how much of a usual MBG portion the student consumed, with six response categories: "*Biasanya tidak dimakan*" (usually not eaten), "*Kurang dari* ¼" (less than one quarter), "*Sekitar* ¼–½" (approximately one quarter to one half), "*Sekitar* ½–¾" (approximately one half to three quarters), "*Lebih dari* ¾" (more than three quarters), and "*Seluruhnya/hampir seluruhnya*" (all or almost all). AT05 recorded how many of the five most recent MBG meals were finished or almost finished, with responses ranging from 0 to 5. The binary endpoint was defined as follows:

$$Y_i = 1\left(\mathrm{AT04}_i \in \left\{\text{"Biasanya tidak dimakan"}, \text{"Kurang dari } \frac{1}{4}\text{"}\right\} \wedge \mathrm{AT05}_i \leq 1\right) \quad (1)$$

A respondent was classified as having persistent low meal utilization ($Y_i = 1$) only when both conditions were met: the usual portion consumed was either not eaten or less than one quarter, and no more than one of the five most recent meals was finished or almost finished. All other respondents were assigned to the comparison class ($Y_i = 0$). This included respondents who reported consuming at least one quarter of the usual portion, as well as those in the two lowest AT04 categories who had finished or almost finished at least two of the five most recent meals. Combining usual portion consumption with recent meal-completion frequency was intended to capture a recurring low-consumption pattern rather than a single low-portion response or an isolated unfinished meal. This endpoint is an operational behavioral definition, not a validated nutritional, clinical, or program-performance threshold. Therefore, membership in the comparison class should not be interpreted as evidence of nutritional adequacy.

The analysis used 21 respondent-reported predictors covering three domains: respondent and school context (province, school level, grade, age, and sex); household characteristics (AH01–AH07, covering guardians' education and occupation, household income range, household size, and social-assistance receipt); and MBG implementation and behavioral context (AT01–AT03 and AT08–AT13). AT04–AT07 were excluded because they define, or are closely related to, the endpoint. Preprocessing was performed using training data only for each model fit. Numeric predictors were imputed using the median, with additional indicators for missing values. Categorical predictors were assigned a constant missing category and then one-hot encoded, with unseen categories ignored. No scaling, rare-category collapsing, class weighting, or resampling was applied. Explicit "do not know" and refusal responses were treated as observed categories.

3.3. Comparative Prediction Models

Three model families were included in the comparative evaluation: Logistic Regression (LR), Random Forest (RF), and XGBoost. The selection was guided by their use in related prediction studies involving health, nutrition, and structured data [6], [19], [24]. LR served as the baseline model, while RF and XGBoost represented alternative approaches to the same prediction task [25], [26]. The candidate configurations and corresponding parameter settings for all three model families are summarized in Table 1. The comparison was not intended to cover all available classification algorithms, but rather to examine the consistency of predictive performance and subsequent analytical findings across these selected model families. Support vector machines and neural networks, for example, were outside the scope of this evaluation.

Table 1. Candidate configurations and selected settings for the three model families

Model	Candidate/search space	Fixed settings	Selected configuration in nested evaluation	Full-data/downstream configuration
Logistic Regression	Inverse regularization strength C = 0.001, 0.01, 0.1, 1, 10 or 100 (6 candidates, LR-C1 to LR-C6)	penalty = L2; solver = lbfgs; max_iter = 5000	LR-C3: C = 0.1 (outer folds 1 to 5)	LR-C3: C = 0.1
Random Forest	Full grid of 12 candidates (RF-C1 to RF-C12): n_estimators = 300 or 500; max_depth = 8, 16 or None (unlimited); min_samples_leaf = 1 or 5	max_features = sqrt	RF-C5: n_estimators = 500, max_depth = 8, min_samples_leaf = 1 (outer folds 1, 2, 3 and 5)	Random Forest
XGBoost	12 predefined candidates (XGB-C1 to XGB-C12), not a full grid, varying n_estimators, max_depth, learning_rate and min_child_weight (listed below the table)	objective = binary:logistic; subsample = 0.8; colsample_bytree = 0.8; reg_lambda = 1; reg_alpha = 0; tree_method = hist; eval_metric = logloss; no early stopping	XGB-C9: (300, 4, 0.03, 10) (outer fold 1)	XGBoost

Note: RF-C5 corresponds to n_estimators = 500, max_depth = 8, and min_samples_leaf = 1. XGBoost configurations are represented as (n_estimators, max_depth, learning_rate, min_child_weight): C1 = (300, 2, 0.03, 5); C2 = (300, 3, 0.03, 5); C3 = (500, 2, 0.03, 10); C4 = (500, 3, 0.03, 10); C5 = (300, 2, 0.07, 5); C6 = (300, 3, 0.07, 5); C7 = (500, 2, 0.07, 10); C8 = (500, 3, 0.07, 10); C9 = (300, 4, 0.03, 10); C10 = (500, 4, 0.03, 10); C11 = (300, 4, 0.07, 10); C12 = (500, 4, 0.07, 10).

All three model families were trained using the same preprocessed feature set and configured to generate predicted probabilities for the binary endpoint. LR and RF were implemented using scikit-learn [44], whereas XGBoost was implemented using the XGBoost package [26]. The random seed was fixed at 20260916. LR and RF were evaluated using complete candidate grids comprising six and twelve configurations, respectively. For XGBoost, twelve configurations were predefined by varying the number of estimators, maximum tree depth, learning rate, and minimum child weight, rather than evaluating the full 24-combination grid that could be formed from the selected parameter values. Parameters not explicitly listed in Table 1 were left at their respective library defaults. Computational settings, including n_jobs = −1 for RF and n_jobs = 4 with device = "cuda" for XGBoost, were used for execution and did not affect the candidate search space.

For each model family, candidate configurations were evaluated within the nested validation procedure described in Section 3.4. The configurations selected during nested evaluation were determined independently within each outer training set and could therefore differ across outer folds. Table 1 also reports the selected configurations and distinguishes them from those used in the subsequent full-data analyses. Models meeting the predefined predictive competitiveness criteria (Fig. 1, stage 3) were retained for the model-choice robustness and province-holdout analyses, whereas the remaining models served as benchmark comparators. A reference model was subsequently designated for downstream interpretation.

3.4. School-Disjoint Model Training and Evaluation

Folds. Respondents from the same school share a setting, so respondent-level random splits could place one school on both sides of a split. In school-disjoint (school-grouped) cross-validation, each school falls in exactly one fold, so no school contributes to both the training and the evaluation data of any split; grouped validation is recommended when data have a dependence structure [45]. The design is nested, with five outer folds and, within each outer training set, four inner folds. The folds were generated once by school-grouped splitting (GroupKFold) in the original analysis environment and reused unchanged for all families and robustness layers, because GroupKFold tie-breaking differs between scikit-learn 1.7.2 and 1.9.0 and regeneration would reassign schools. Analyses ran in a pinned environment (Python 3.13.3, scikit-learn 1.7.2, XGBoost 3.4.1, shap 0.52.0). Selection and prediction. Within each outer training set, each candidate configuration is fitted on the inner training folds and scored on the pooled inner validation predictions by PR-AUC, Brier score, and log loss. The configuration with the highest PR-AUC is selected, with ties resolved by lower Brier score, lower log loss, and candidate order. It is refitted on the whole outer training set and predicts the held-out outer fold. Repeating selection inside each outer training set separates the data used to choose a configuration from the data used to evaluate it [29]. The five held-out predictions form one out-of-fold prediction per respondent.

Metrics. Discrimination is summarized by the area under the precision-recall curve (PR-AUC; computed as average precision, the non-interpolated scikit-learn implementation) and the area under the receiver operating characteristic curve (ROC-AUC). Because the precision-recall baseline equals the outcome prevalence [46], PR-AUC is interpreted relative to prevalence, as PR-AUC minus prevalence and as a ratio. Probability accuracy is summarized by the Brier score, log loss, and the Brier skill score against a null prediction equal to the outcome prevalence of the respective outer training fold [47]. The utility assessment considered three criteria for each family: a PR-AUC-to-prevalence ratio of at least 1.20, a Brier skill score above 0, and a lower 95% bootstrap bound for PR-AUC minus prevalence above 0.

Calibration [9] was summarized by reliability over ten equal-count bins and by the calibration intercept and slope from an unpenalized logistic recalibration of the endpoint on the logit of the predicted probability, both estimated jointly (the intercept is therefore not calibration in the large). Expected calibration error (ECE) [48] is a secondary descriptive measure and is not used for selection. To compare the models as fitted, no recalibration was applied; calibration results therefore describe the raw probability outputs of the selected configurations. Uncertainty. A school-cluster bootstrap [49] resampled schools with replacement (1,000 replicates, seed 20260916), retained every respondent of a drawn school, and recomputed each metric on the resampled pooled predictions; intervals are the 2.5th and 97.5th percentiles, and differences between models are paired because the same resampled schools were used for all models in a comparison. Fitted models and predictions are held fixed, so intervals are conditional on the fitted models and exclude variability from model fitting and configuration selection. No multiplicity adjustment was applied, and intervals are interpreted descriptively.

3.5. Model-Choice Robustness Analysis

This analysis examines whether the principal downstream outputs depend on the choice among the models retained for downstream evaluation (Fig. 1, stage 4). Each retained model received its own operational threshold (Section 3.7, applied to its own out-of-fold predictions), its own full-data fit for scenario probabilities, and the same scenario rules. The outputs compared were: (i) out-of-fold predicted probabilities (Spearman correlation as the primary measure, Pearson correlation and absolute differences); (ii) thresholds with bootstrap

intervals; (iii) high-risk classification at each model's own fold-specific thresholds (proportion classified high-risk, agreement, Cohen's kappa, and discordant respondents); (iv) source-variable predictive attribution ranks (Spearman correlation across variables, overlap of the top three, five and ten variables, and ranks of the scenario-mutable variables), compared by rank only because attribution magnitudes from different explainers are not comparable; (v) cross-model scenario agreement (scenario yield, respondent-level agreement in crossing outcome, and agreement of selected actions); and (vi) the direction and size of the predicted change for each AT10 transition. Paired school-cluster intervals (Section 3.4) accompany these measures where a sampling distribution is meaningful; scenario intervals are conditional on the fitted full-data models. No numeric pass or fail cut-off was applied, and the comparison of downstream outputs is descriptive.

3.6. Geographic Robustness Evaluation

Province-holdout (leave-one-province-out) evaluation examines internal geographic robustness: predictive performance and calibration when all respondents from a province are absent from model development. Because all held-out provinces come from the same survey, this procedure assesses internal geographic robustness rather than independent external validation [11] and does not support claims of national representativeness or transportability. It follows the logic of internal-external validation by natural units [11], with the blocking rationale for grouped validation [45].

Provinces with at least 100 respondents in the analytic cohort were eligible for holdout; respondents from the other provinces stayed in the training data of every fit and were never held out. Each school belongs to one province. In the primary design (Design A), province is removed from the predictors because a fully held-out province is an unseen category when province is a predictor; in the sensitivity design (Design B), province is retained, and the held-out province is handled by the encoder's unknown-category rule. The A/B comparison is therefore interpreted as a sensitivity analysis of province encoding rather than as a component-wise ablation. For each eligible province and design, all its respondents were held out, and the retained models were trained on all other respondents, with the configuration chosen inside the training data by school-grouped four-fold cross-validation, using the candidate sets and selection rule of Sections 3.3 and 3.4 and preprocessing fitted on training data only. Each holdout has a different training set, so its inner folds were generated by school-grouped splitting in the pinned environment.

For each held-out province, PR-AUC (with the province prevalence), ROC-AUC, and Brier score were computed. For the pooled province-holdout predictions, PR-AUC, ROC-AUC, Brier score, log loss, calibration intercept and slope, and ECE were computed, together with the median, interquartile range and range across provinces. The pooled estimate was compared with the school-grouped estimate for the same respondents. Uncertainty came from a paired school-cluster bootstrap (1,000 replicates; schools resampled within province for province-level intervals and across provinces for pooled intervals). Calibration quantities for provinces with fewer than 100 events or fewer than 100 non-events were flagged as of limited reliability, applying the minimum suggested for external validation samples [50]. Between-province variation was compared with the sampling variation implied by the within-province bootstrap.

3.7. Reference-Model Interpretation and Operational Thresholding

The designated reference model was Logistic Regression; its coefficients and closed-form Linear SHAP values support direct inspection of the fitted predictive structure. Coefficients and predictive attribution. Coefficients (log-odds scale) were mapped from encoded columns to source variables and categories through the fitted preprocessor. They describe the structure of the fitted predictive model, that is, conditional associations given the other predictors, not effects of survey items. For each outer fold, the model refitted on the training schools yielded interventional Linear SHAP values under a feature-independence assumption [30] for the held-out schools only, $\varphi_{ij} = \beta_j\left(x_{ij} - \hat{x}_j\right)$ on the log-odds scale, where $\hat{x}_j$ is the mean of encoded column j in the outer training fold; shap.LinearExplainer with an independent masker reproduces these values. "Interventional" refers to the use of the background distribution, not to a causal intervention. Encoded attributions were grouped by source

variable V, summing within respondent with signs preserved and taking the absolute value only afterwards, as in Eq. (2):

$$\Phi_{iV} = \sum_{j \in J(V)} \varphi_{ij}, \qquad I_V = \frac{1}{N} \sum_{i=1}^{N} |\Phi_{iV}|. \tag{2}$$

Here $J(V)$ is the set of encoded columns belonging to source variable V, and N is the number of respondents (held-out predictions pooled over the five outer folds).

Source variables were ranked by I_V. Rank stability was assessed with the attributions fixed, by a school-cluster bootstrap of respondents (1,000 replicates) that recomputed the ranks and the proportion of replicates in which each variable ranked in the top 3, 5, and 10. Where a retained model is a tree ensemble, Tree SHAP [31] values on the raw log-odds scale were computed for the held-out schools of each outer fold and grouped identically. SHAP values are model-dependent attributions, not causal importance; Shapley-based importance has known limits [32].

Operational threshold. High-risk classification uses the F2 criterion, the case β=2 of F_β [51], which weights recall above precision, as is appropriate for a classification rule intended for screening purposes: $F_2 = 5PR/(4P + R)$ for precision P and recall R. Fold-specific threshold selection uses out-of-fold predictions. For outer fold k, the selection pool is the pooled out-of-fold predictions and observed endpoints of the respondents in the other four folds; the observations of fold k were excluded from threshold selection. These pool predictions were generated by the original outer-fold models and may therefore originate from fits that included observations from fold k. The candidates are the unique predicted probabilities in the pool; the F_2 maximizer is chosen, with ties resolved by higher recall, then higher precision, then the higher threshold. The fold-specific threshold τ_k is applied once to fold k, whose respondents are classified high-risk if $p_i \geq \tau_k$. The reference threshold is the median of the five-fold-specific thresholds and was not reoptimized on the pooled predictions, as in Eq. (3):

$$\hat{\tau} = \text{median}(\tau_1, \ldots, \tau_5), \qquad \tau_k = \arg\max_{\tau \in \mathcal{C}_k} F_2^{(-k)}(\tau). \tag{3}$$

Here $\mathcal{C}_k$ is the candidate set and $F_2^{(-k)}$ is computed on the selection pool of fold k. The out-of-fold high-risk rate at fold-specific thresholds is the share of respondents classified high-risk with their own fold's threshold. Threshold uncertainty was quantified with the school-cluster bootstrap by reselecting the F_2-maximizing threshold on each resample of the pooled out-of-fold predictions without refitting; this interval describes the pooled threshold and does not repeat the fold-specific procedure. The threshold is an operational classification threshold, not a diagnostic criterion, and is an input to, not an output of, the scenario analysis.

3.8. Rule-Constrained Minimal-Action Scenario and Objective-Alignment Analysis

Status of the analysis. The reference-model scenario analysis asks whether the designated reference model's predicted probability for a respondent classified as high-risk falls below the reference threshold when a few scenario-mutable survey responses are set to other permitted answers. It is a model-based, hypothetical analysis of model inputs and predicted probabilities. Because these probabilities come from a model fitted to the full analytic dataset, the scenario outputs are in-sample and hypothetical. They should not be interpreted as estimates of real-world intervention effects, because the consequences of an action depend on causal structure that the model does not represent [40].

Model and eligibility. Scenario baselines and probabilities come from the reference model refitted on all analytic data with the same preprocessing specification, using the configuration chosen by applying the selection key of Section 3.4 to the pooled school-grouped out-of-fold predictions of the candidate configurations. They are in-sample and serve only as hypothetical model outputs. A respondent is scenario-eligible if the full-data predicted probability is at or above the reference threshold and at least one predefined transition applies; the observed endpoint plays no role.

Variables and allowed changes. Four survey items are scenario-mutable: AT02 (days of scheduled MBG not received in the last four weeks), AT03 (usual timing of receipt), AT10 (perceived portion size) and AT12 (menu variation). All other predictors, including the

diagnostic items AT08, AT09 and AT11, remain fixed. Allowed changes follow a predefined, directionally constrained transition map that moves an answer toward the item's target state: fewer days of missed deliveries (AT02), receipt at an appropriate time (AT03), a sufficient portion (AT10) and a more varied menu (AT12). For AT10, the destination is "*Cukup*" ("sufficient") from both under-sized and over-sized states. No transition to or from "*Tidak tahu*" ("do not know"), reversed transition, or unlisted transition is permitted.

Search and selection. An action changes m=1 to 4 scenario-mutable variables, each once. For each eligible respondent, k increases until at least one action yields a scenario threshold crossing, that is, a scenario probability $p_i^a < \hat{\tau}$. Risk reduction is $p_i - p_i^a$, with changes within 10^{-12} of zero treated as neutral. Among the crossing actions at the smallest m, the selected action minimizes, in order, negative risk reduction, total ordinal category distance, variable order, and category order. A respondent with no crossing at m≤4 is recorded as having no scenario threshold crossing.

Empirical-support qualification. The support of the selected action is the number of respondents in the analytic cohort whose values on the four scenario-mutable variables equal the post-action profile (the exact tuple): 30 or more is high support, 10 to 29 moderate, 1 to 9 low, and 0 unobserved. A scenario is support-qualified if it has a crossing and high exact-tuple support; other crossings are exploratory. Support describes how often a profile occurs in the data; it is not evidence of feasibility or of an effect.

Objective alignment. The endpoint measures behavioral utilization, not nutritional adequacy, and nutrient intake was not observed, so a predicted reduction in the risk of low utilization does not imply improved nutritional adequacy. For each AT10 transition toward "Cukup", the change in predicted probability was summarized among eligible respondents to whom the transition applies by direction (share with lower, higher or unchanged probability), mean change and share crossing the threshold, for each retained model, to examine whether the direction of predicted change depends on the starting state (under-sized or over-sized) for a common destination.

4. Results and Discussion

4.1. Study Cohort and Geographic Coverage

Of the 11,955 survey submissions, 8,964 respondents reported currently receiving MBG meals. The final analytic cohort comprised 8,802 respondents from 1,337 schools who had an observed endpoint and a valid school identifier. Persistent low meal utilization was identified in 2,550 respondents (28.97%). Under the school-disjoint validation design, each of the five outer folds held out 267 or 268 schools, corresponding to 1,760 or 1,761 respondents.

The cohort covered 38 province groups, with no school identifier appearing in more than one province. However, geographic representation was uneven. Banten contributed 2,733 respondents (31.05% of the cohort) from 193 schools, while the three largest provinces together accounted for 47.0% of respondents. Seventeen provinces had at least 100 respondents (range: 103–2,733) and together contributed 8,251 respondents (93.74% of the cohort) from 1,173 schools. These provinces contained 2,422 endpoint cases (29.35%) and formed the province-holdout evaluation set. The remaining 21 provinces contributed 551 respondents (6.26%) and were included in model training but not in province-level evaluation. Endpoint prevalence across the 17 eligible provinces ranged from 19.2% to 42.6%. Thus, although the province-holdout evaluation covered most of the analytic cohort, it did not provide direct geographic validation for all 38 province groups.

Regarding respondent characteristics, 6,712 respondents (76.26%) were students, 1,337 (15.19%) were students completing the questionnaire together with a parent or guardian, and 753 (8.55%) were parents or guardians. Most respondents were from general or Islamic senior secondary schools (SMA/MA; 5,735, 65.16%) and vocational secondary schools (SMK; 1,526, 17.34%), with the remainder from junior secondary, primary, other, or special schools. The median number of respondents per school was 3 (interquartile range: 1–7). A total of 433 schools contributed only one respondent, while the ten largest schools accounted for 10.91% of the cohort. These figures indicate substantial variation in school-level sample sizes, which further supports the use of school-disjoint validation rather than respondent-wise random splitting. Nevertheless, the cohort represents a voluntary-response sample and should not be interpreted as representative of the student population in any province or in Indonesia.

4.2. Comparative Predictive Performance

Pooled out-of-fold performance of the three model families under the frozen school-disjoint nested protocol is summarized in Table 2 and Figure 2. All three models achieved similar discrimination, with PR-AUC (average precision) values of 0.7270 for LR (95% CI: 0.7067–0.7480), 0.7213 for RF (0.7002–0.7426), and 0.7277 for XGBoost (0.7073–0.7486), against an endpoint prevalence of 0.2897. Their PR-AUC values were approximately 2.5 times the prevalence baseline, and all three model families met the predefined predictive utility criteria.

Table 2. Pooled out-of-fold predictive performance and calibration of the three model families under the frozen school-disjoint nested protocol (8,802 respondents)

Model	PR-AUC (95% CI)	ROC-AUC	PR-AUC to prevalence	Brier skill score	Calibration slope	Calibration intercept	ECE
Logistic Regression	0.7270 (0.7067 to 0.7480)	0.8787	2.51	0.3856	1.070	0.044	0.0204
Random Forest	0.7213 (0.7002 to 0.7426)	0.8762	2.49	0.3141	2.203	0.829	0.1094
XGBoost	0.7277 (0.7073 to 0.7486)	0.8805	2.51	0.3893	1.113	0.066	0.0205

Note: Endpoint prevalence 0.2897 (PR-AUC baseline). ECE, expected calibration error. Intervals are 95% school-cluster bootstrap intervals; calibration is on the raw probability scale (no recalibration).

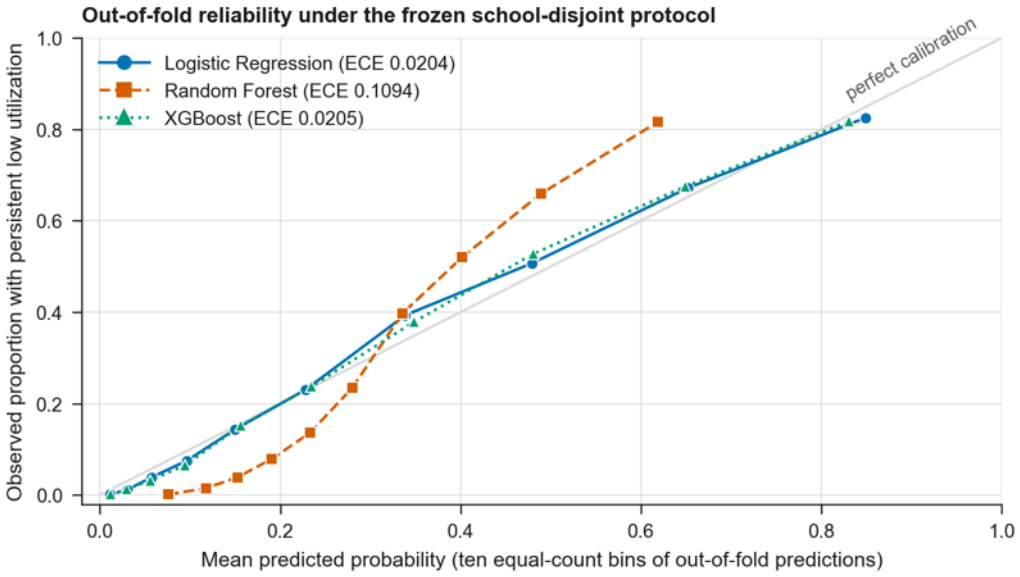


Figure 2. Reliability of out-of-fold predicted probabilities for the three model families (8,802 respondents; raw probabilities, no recalibration)

In paired school-cluster bootstrap comparisons, the PR-AUC difference between XGBoost and LR was small and its confidence interval included zero (+0.0007; 95% CI: −0.0035 to +0.0045). XGBoost showed small advantages over LR in ROC-AUC (+0.0017; 95% CI: +0.0002 to +0.0033) and log loss (−0.0025; 95% CI: −0.0047 to −0.0005), while the Brier score difference was −0.0008 (95% CI: −0.0016 to +0.0000). These differences indicate a marginal advantage for XGBoost in some metrics, but do not substantially distinguish its overall predictive performance from LR.

RF showed a different pattern. Its PR-AUC differences relative to LR and XGBoost included zero (XGBoost minus RF: +0.0064, 95% CI: −0.0009 to +0.0131; LR minus RF: +0.0057, 95% CI: −0.0013 to +0.0123). However, its Brier score and log loss were higher than those of both models, with paired differences excluding zero for the reported comparisons (Brier score differences: −0.0148 for LR minus RF and −0.0155 for XGBoost minus RF; log loss differences: −0.0500 and −0.0525, respectively). This distinction suggests that

RF retained broadly comparable ranking ability but produced less reliable probability estimates in the fitted configuration.

The calibration results further illustrate this difference (Figure 2). LR and XGBoost showed similar calibration slopes (1.070 and 1.113) and expected calibration errors (ECE: 0.0204 and 0.0205), whereas RF had a substantially higher calibration slope (2.203) and ECE (0.1094). The reliability curves show how predicted probabilities corresponded to observed proportions of persistent low meal utilization across ten equal-count bins. The diagonal represents perfect calibration, and the ECE for each model is reported in the figure legend.

Given their comparable discrimination and calibration, LR and XGBoost were retained for the subsequent model-choice robustness and geographic evaluations, while RF was retained as a benchmark comparator. This selection provides a more meaningful basis for examining whether two competitive models produce consistent downstream analytical findings, rather than comparing a competitive model with one that has substantially weaker probability calibration. Since no formal equivalence test was performed, the performance of LR and XGBoost is described as similar rather than statistically equivalent. Furthermore, because no recalibration was applied, the calibration findings for RF should be interpreted as specific to the fitted configuration and not generalized to the Random Forest algorithm as a whole.

4.3. Downstream Model-Choice Robustness

The two models showed substantial agreement across several downstream analytical outputs, although differences remained in individual classifications and scenario outcomes. Their agreement was particularly strong in probability ranking and predictive attribution, whereas threshold-based classification and scenario analysis revealed more noticeable model-dependent differences.

The out-of-fold predicted probabilities from LR and XGBoost were highly correlated (Spearman's 0.9874, 95% CI 0.9863-0.9883), with a median absolute difference of 0.0187. This indicates that the models produced very similar probability rankings, although their predicted probabilities were not identical at the respondent level. The median of the fivefold-specific operational thresholds was 0.1695 for LR (95% interval 0.1404–0.2024) and 0.1558 for XGBoost (0.1403–0.2002), with each estimate falling within the other's interval. When these model-specific thresholds were applied to their respective out-of-fold predictions, 52.76% of respondents were classified as high-risk by LR, compared with 54.26% by XGBoost, a difference of 1.50 percentage points (95% CI 0.95–2.08). Overall classification agreement was 94.59% (95% CI 94.14–95.11; Cohen's kappa 0.891, 95% CI 0.882–0.902). Nevertheless, 476 respondents (5.41%) received different classifications: 304 were classified as high-risk only by XGBoost and 172 only by LR. Thus, the high correlation in predicted probabilities translated into substantial, but not complete, agreement in individual-level classification.

A similar pattern of agreement was observed in predictive attribution. The source-variable rankings obtained from Linear SHAP for LR and Tree SHAP for XGBoost shared the same three leading variables (AT11, AT10, and AT09) and all five leading variables, with nine variables in common among the top ten. The bootstrap 95% interval for the top-five overlap was 4–5 variables, while the Spearman correlation across all 21 source variables was 0.886 (95% CI 0.860–0.895). AT10, which was subsequently examined in the scenario analysis, ranked second under LR and third under XGBoost. The remaining scenario-related variables, AT03, AT12, and AT02, ranked 10th, 14th, and 15th under LR, compared with 11th, 15th, and 17th under XGBoost, respectively. These results indicate that the two models emphasized broadly similar predictive features, even though they did not always assign the same respondents to the high-risk category. Agreement in cohort-level feature rankings therefore did not necessarily imply agreement in individual-level predictions.

Differences were also evident in the full-data scenario analysis. Under XGBoost, 4,852 respondents (55.12% of the cohort) were classified as high-risk, of whom 4,622 were eligible for scenario analysis; 308 eligible respondents (6.66%) experienced a threshold crossing, including 307 with one changed variable and one with two. Under LR, 4,307 respondents were eligible, with 382 threshold crossings (8.87%). Among the 4,236 respondents eligible under both models, crossing outcomes agreed for 96.22% (95% CI 95.57–96.81; Cohen's kappa 0.704, 95% CI 0.659–0.744). Of the 209 respondents who experienced a crossing under both models, 205 (98.09%) had the same selected action. AT10 was involved in 292 of the 308 XGBoost crossings (94.81%) and 313 of the 382 LR crossings (81.94%). Eligibility was

determined separately using each model's full-data predictions and operational threshold, accounting for the different eligible populations. The difference in crossing rates therefore reflects model-dependent scenario outputs as well as differences in the populations eligible for analysis. The analysis does not isolate the contribution of baseline predictions, operational thresholds, and model responses to input changes to the observed difference in scenario yield.

Despite these differences in crossing rates, both models produced consistent directional responses to the four AT10 transitions toward "*Cukup*" (sufficient). Transitions from "*Sangat kurang*" and "Kurang" (too little) increased predicted probabilities, with mean changes of +0.3148 and +0.2131 under LR and +0.3300 and +0.1975 under XGBoost, respectively. Conversely, transitions from "*Agak berlebihan*" and "*Sangat berlebihan*" (too much) reduced predicted probabilities, with corresponding mean changes of −0.1684 and −0.2534 under LR and −0.1547 and −0.2581 under XGBoost. Every respondent to whom a given transition applied exhibited the same direction of change within each model. However, consistency in direction did not imply identical response magnitudes or scenario outcomes.

The results reveal that model agreement depends on the analytical level being examined. LR and XGBoost produced highly similar probability rankings and predictive attribution patterns, but retained differences in individual classifications and scenario yields. This suggests that comparable aggregate performance and feature rankings should not, by themselves, be taken as evidence that competing models are interchangeable in downstream analyses.

4.4. Geographic Robustness

Pooled performance under province holdout remained closely aligned with school-grouped performance for both models. In the primary design (Design A, province removed from the predictors), pooled PR-AUC was 0.7285 for XGBoost (95% CI 0.7063–0.7519) and 0.7288 for LR (95% CI 0.7064–0.7526), with an XGBoost-minus-LR difference of −0.0003 (95% CI −0.0036 to +0.0028). These results were consistent with the school-grouped estimates for the same 8,251 respondents across the 17 eligible provinces (0.7293 for XGBoost and 0.7290 for LR), indicating that withholding entire provinces did not materially alter the aggregate comparison. Pooled ROC-AUC and calibration measures also showed similar performance between the two models.

Although province-level performance varied across held-out provinces (Figure 3), PR-AUC exceeded the corresponding endpoint prevalence in every province. The median province-level PR-AUC was 0.7369 for XGBoost and 0.7421 for LR. The heterogeneity in province-level PR-AUC relative to within-province bootstrap sampling variance was limited (I^2 = 13.5% for XGBoost and 3.9% for LR). Importantly, the median of the paired province-level PR-AUC differences (XGBoost minus LR) was +0.0009 (range −0.0101 to +0.0262). XGBoost achieved higher PR-AUC in 9 of the 17 provinces, but none of the province-level 95% bootstrap intervals for the paired differences excluded zero. Thus, neither model demonstrated clear and consistent superiority across the evaluated provinces.

In the sensitivity design retaining province as a predictor (Design B), pooled PR-AUC was 0.7296 for XGBoost and 0.7287 for LR. Compared with Design A, the changes were small (+0.0010 for XGBoost and −0.0001 for LR), suggesting that including province information did not materially affect aggregate discrimination. However, province-level calibration estimates, including calibration intercept, slope, and expected calibration error, were flagged as having limited reliability in 9 of the 17 provinces because these provinces had fewer than 100 events or fewer than 100 non-events.

Overall, province-holdout evaluation preserved the comparative LR/XGBoost conclusion, with both models maintaining closely aligned aggregate discrimination when entire provinces were withheld from model development. These findings support internal geographic robustness of the aggregate model comparison. Nevertheless, they do not establish invariance of individual predictions, national representativeness, transportability to unobserved geographic settings, or independent external validity.

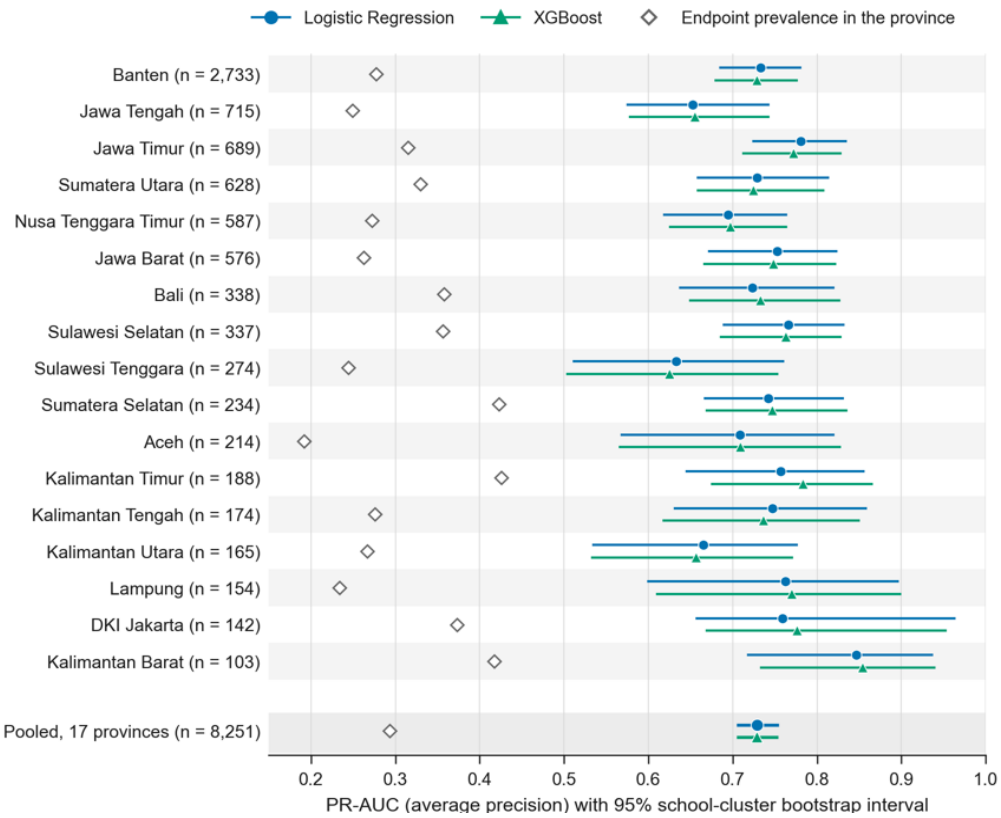


Figure 3. Province-level PR-AUC under province holdout (Design A) for the 17 eligible provinces, ordered by number of respondents (n)

4.5. Reference-Model Predictive Interpretation

LR was designated as the reference model for downstream interpretation. This designation considered its competitive predictive performance and calibration, consistency with XGBoost in the downstream comparisons, geographic robustness, and simpler additive structure with transparent coefficient-based interpretation. This choice represents a documented judgment based on the combined evidence rather than a claim of statistical superiority over XGBoost.

The source-grouped Linear SHAP analysis of the reference model, computed using held-out schools in the out-of-fold context, identified AT11 (taste liking; mean absolute grouped contribution of 0.9505 on the log-odds scale), AT10 (perceived portion size; 0.6015), AT09 (frequency of still buying other main food; 0.5570), AT08 (frequency of also bringing food from home; 0.2433), and AH05 (household income range; 0.1199) as the five highest-ranked variables (Figure 4(a)). The stability of these rankings was further examined through 1,000 school-cluster bootstrap replicates. The top three variables consistently formed the same set in all replicates (100%), with AT11, AT10, AT09, and AT08 retaining ranks 1, 2, 3, and 4, respectively. The fifth position was less stable: AH05 appeared in the top-five set in 84.8% of replicates, while AH01 (father or male guardian's education) appeared in 15.2%. Among the scenario-mutable variables, AT10 ranked second, AT03 tenth, AT12 fourteenth, and AT02 fifteenth; notably, AT03 appeared in the top ten in 98.3% of replicates. Figure 4(a) also presents the 95% school-cluster bootstrap intervals, highlights scenario-mutable variables using hatched bars, and reports the corresponding XGBoost ranks using Tree SHAP. Only ranks are shown for XGBoost because attribution magnitudes are not directly comparable across the two explainers.

The fitted LR coefficients provide additional detail on how individual response categories contributed to the model predictions. The largest category-level coefficient magnitudes were observed for AT11, ranging from −1.643 for "*Sangat suka*" (liked very much) to +1.869 for "*Sangat tidak suka*" (disliked very much); AT10, ranging from −1.439 for "*Sangat kurang*" to +1.436 for "*Sangat berlebihan*"; (with −0.007 for "*Cukup*") and AT09, ranging from −1.022 for "*Tidak pernah*" (never) to +1.056 for "Selalu" (always). These coefficients complement the

global attribution results by showing how different response categories contribute to the predicted log-odds, while remaining predictive associations rather than causal effects.

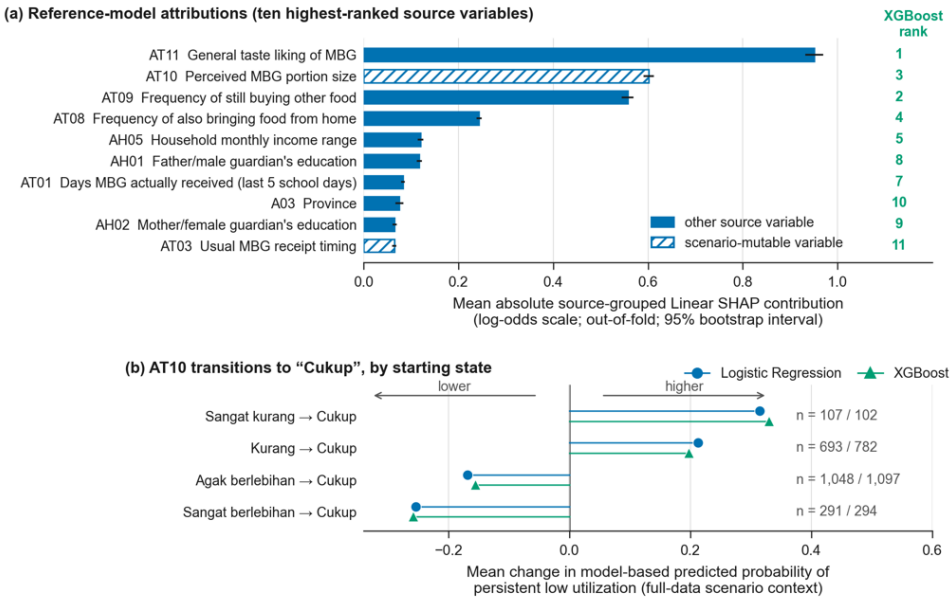


Figure 4. (**a**) Source-grouped Linear SHAP attribution of the reference model for the ten highest-ranked source variables. (**b**) Mean change in predicted probability of persistent low utilization for each AT10 transition to "*Cukup*", by starting state

4.6. Operational Threshold, Scenario, and Objective-Alignment Results

Reference threshold. The reference threshold, the median of the fivefold-specific F2 thresholds (0.1504, 0.1695, 0.1505, 0.1717 and 0.1717), was 0.1695 (bootstrap 95% interval 0.1404 to 0.2024). In the out-of-fold context, 4,644 of 8,802 respondents (52.76%) were classified as high-risk with their fold-specific thresholds. At these thresholds, the classification identified 2,352 of the 2,550 respondents with the endpoint (sensitivity 0.9224) and 3,960 of the 6,252 without it (specificity 0.6334); precision was 0.5065, and F2 was 0.7922.

Scenario yield (full-data scenario context). The scenario results in this section come from models fitted on all analytic data; they are in-sample, hypothetical model outputs. With the full-data reference model, 4,525 respondents (51.41% of the 8,802-respondent cohort) had a predicted probability at or above the reference threshold; this full-data scenario high-risk population is a different quantity from the out-of-fold high-risk rate above. Of the 4,525, 4,307 (95.18%) had at least one applicable transition and were scenario-eligible. Among the 4,307, 382 (8.87%) had a scenario threshold crossing under at least one admissible minimal-action scenario: 357 (8.29%) with one changed variable and 25 (0.58%) with two, and none with three or four. No crossing was found for the other 3,925 (91.13%). Across the 382 crossings, the median baseline predicted probability was 0.2484, the median selected-scenario probability was 0.1336, and the median respondent-level reduction in predicted probability was 0.1368 (a median of respondent-level differences).

Empirical support and action patterns. Of the 382 crossings, 292 (76.44%) had exact tuple support of at least 30 respondents and were support-qualified, and 90 (23.56%) were exploratory (69 with support of 10 to 29 and 21 with support of 1 to 9). AT10 was the only changed variable in 294 of the 382 crossings; AT12 was involved in 78 (20.42%), AT02 in 16 (4.19%), and AT03 in none.

Objective alignment. Among LR's scenario-eligible respondents to whom each AT10 transition applied, changing AT10 from "*Sangat kurang*" (n = 107) or "Kurang" (n = 693) to "*Cukup*" raised the predicted probability for every respondent (no crossings), whereas changing it from "*Agak berlebihan*" (n = 1,048) or "*Sangat berlebihan*" (n = 291) to "*Cukup*" lowered it for every respondent (20.99% and 25.43% of these respondents crossed the reference threshold). The direction of the predicted change thus depended on the starting AT10 state, as it did for XGBoost (Fig. 4(b)). These in-sample predictions describe model behavior and should not be interpreted as evidence that changing perceived portion size would change

actual utilization. In Fig. 4(b), the changes are in-sample, are shown for each model's own eligible respondents (n = Logistic Regression / XGBoost), and carry no causal interpretation.

4.7. Discussion

The results indicate that model-choice robustness is dependent on the analytical layer being examined. Logistic Regression (LR) and XGBoost exhibit closely aligned aggregate predictive performance, calibration characteristics, leading predictive attribution, and pooled geographic performance. However, this similarity does not translate into complete agreement in individual-level classifications and model-based scenario outputs. These findings suggest that comparable aggregate performance should not automatically be interpreted as evidence that competing models are interchangeable in downstream analyses.

At the individual level, the observed disagreement between LR and XGBoost demonstrates that highly similar probability rankings can still produce different classifications when model-specific thresholds are applied. This distinction is important because aggregate discrimination metrics do not fully capture the extent to which model selection may affect individual-level decisions. Consequently, classification agreement provides complementary information for assessing model-choice robustness rather than serving as a replacement for conventional predictive performance metrics.

The predictive attribution results provide further evidence of cross-model consistency, particularly in the leading variables. Both models identify the same dominant predictors, with substantial overlap among their highest-ranked variables. This suggests that the main predictive signals are relatively stable across a linear additive model and a boosted tree model capable of representing nonlinear relationships. Nevertheless, agreement in global attribution does not necessarily imply identical individual predictions or identical interpretations of all variables. Furthermore, predictive attribution should not be interpreted as evidence of causal influence[32], [34]. In this study, variables such as AT08, AT09, and AT11 are therefore better understood as diagnostic indicators associated with the prediction task rather than directly modifiable determinants.

The geographic evaluation also requires a distinction between pooled performance and province-level reliability. The similar pooled PR-AUC values indicate that the relative predictive performance of LR and XGBoost remains stable under province-holdout evaluation. However, variations across individual provinces and the limited calibration reliability observed in some province-level evaluations suggest that pooled results may conceal local differences. This is particularly relevant when interpreting model performance across geographically heterogeneous populations [11], [27], [28]. In addition, the geographic evaluation should be interpreted within the coverage of the available sample: the uneven respondent distribution, including the concentration of respondents in Banten, and the exclusion of smaller provinces from the holdout evaluation limit the extent to which these findings can establish broader geographic generalizability.

Table 3 summarizes the main findings across the different analytical layers and highlights the distinction between aggregate model consistency and downstream disagreement. Taken together, the results show that model-choice robustness is not a single property that can be established through one performance metric. Instead, it needs to be assessed across predictive performance, individual classifications, predictive attribution, geographic evaluation, and scenario-based outputs.

The scenario analysis further illustrates why this distinction matters. Although the two models produce broadly consistent patterns, differences in predicted probabilities and model-specific thresholds result in different eligibility and scenario outcomes. The observed AT10 pattern, in which predicted probabilities move toward the reference category of "*Cukup*" from both smaller and larger portion categories, is consistent across the two models. However, this pattern represents a predictive association rather than evidence that changing portion sizes would cause improved meal utilization. The scenario analysis is hypothetical and based on model predictions; it does not estimate the actual effects of interventions involving portion size, menu composition, meal timing, or delivery arrangements. Establishing such effects would require an appropriate causal design and additional evidence [41]. The apparent consistency across models therefore strengthens confidence in the observed predictive pattern, but does not establish its causal validity or practical effectiveness.

Table 3. Synthesis of the main findings: conclusions shared by Logistic Regression and XGBoost and conclusions that depended on the model or on the setting

Evidence layer	Where the models agreed	Where they differed, or the evidence was limited
Discrimination and calibration	PR-AUC 0.7270 (LR) and 0.7277 (XGBoost); similar calibration (slope 1.070 and 1.113)	Small XGBoost edges in ROC-AUC and log loss; RF probabilities less well calibrated (slope 2.203) in the fitted configuration
Probabilities, thresholds and high-risk classification	Out-of-fold probabilities highly correlated (Spearman 0.9874); classification agreement 94.59%	Thresholds 0.1695 (LR) and 0.1558 (XGBoost); 476 respondents (5.41%) classified differently
Predictive attribution	Same three leading variables (AT11, AT10, AT09); nine of the top ten shared	Rank correlation across 21 variables 0.886; fifth rank of LR unstable; AT10 ranked 2 (LR) and 3 (XGBoost)
Scenario analysis, in-sample	Same sign of the AT10 change for all four transitions; no crossing for most eligible respondents; same selected action for 98.09% of shared crossings	Crossings in 382 of 4,307 eligible respondents (8.87%) under LR and 308 of 4,622 (6.66%) under XGBoost; AT12 in 78 and 14 crossings
Province holdout, internal	Pooled PR-AUC 0.7285 (XGBoost) and 0.7288 (LR), close to school-grouped 0.7293 and 0.7290; the interval of the PR-AUC difference included zero in all 17 provinces.	Province-level PR-AUC varied (0.6249 to 0.8541 for XGBoost, 0.6331 to 0.8464 for LR); calibration of limited reliability in 9 of 17 provinces; 21 smaller provinces not evaluated.

Note: LR, Logistic Regression; PR-AUC, area under the precision-recall curve. Scenario values are in-sample, hypothetical model outputs; province holdout is internal geographic robustness, not external validation.

Several limitations should be considered when interpreting these findings. First, the voluntary-response survey design and the absence of population weighting mean that the results describe the observed respondent cohort rather than providing nationally representative estimates of school meal utilization. Second, the analysis is predictive and observational, and therefore cannot establish causal relationships. Third, although the study compares three model families, only LR and XGBoost demonstrate sufficiently competitive performance for the principal downstream comparison. The findings should therefore not be generalized to all possible model families. Fourth, the geographic evaluation assesses internal robustness through province-holdout validation, not independent external validation or temporal stability. Moreover, the sensitivity analyses involving alternative province-encoding designs should not be interpreted as a complete component-level ablation study. Finally, the study's ethics and consent-related limitations, including the absence of a formal ethics review or exemption determination, should be considered in the context of the procedures described in Section 3.2.

Despite these limitations, the findings have practical implications for predictive studies involving heterogeneous populations. When competing models exhibit similar aggregate performance, researchers should examine whether their outputs remain consistent at the individual and downstream analytical levels before treating them as interchangeable. Stable leading predictors may help prioritize further data collection and investigation, whereas model-dependent classifications and scenario outputs warrant additional scrutiny. Any operational or policy recommendations should be supported by independent validation and, where causal effects are claimed, appropriate intervention-based evidence [42]. In this respect, the MBG setting provides an empirical context for examining model-choice robustness, while the methodological implications extend to other predictive applications involving voluntary-response data and geographically heterogeneous populations.

5. Conclusions

Model-choice robustness in this study was layer-specific: conclusions that held at the aggregate level did not necessarily hold for individual classifications and scenario outputs. Similar headline predictive performance therefore does not by itself establish downstream interchangeability, and model comparison should extend to the layers on which model outputs are actually used. The proposed multi-layer framework, which combines school-disjoint model comparison, downstream agreement, predictive attribution, scenario analysis, and geographic robustness, distinguishes conclusions that are stable across competitive models from those that remain model-dependent.

These findings are predictive and model-based and do not establish causal effects; the voluntary-response sample is not nationally representative, and the province holdout provides internal rather than independent external validation. Future work can assess independent external and temporal robustness, test additional competitive models for multiplicity, and combine model-choice robustness with longitudinal, quasi-experimental, experimental or other causally informative designs to test whether consistently identified predictive factors are actionable determinants of meal consumption. Stable explainable-AI findings can help prioritize such questions, but causal inference is required before they are translated into evidence-based recommendations for meal provision or policy.

Author Contributions: Conceptualization: APZ; Methodology: APZ, AA, WHS, NRTN, and SJA; Software: APZ and WHS; Validation: APZ, AA, WHS, NRTN, and SJA; Formal analysis: APZ and WHS; Investigation: APZ, AA, WHS, and MA; Data curation: APZ, WHS, MA, and NRTN; Visualization, APZ and MA; Supervision, APZ and AA; Project administration, APZ; Funding acquisition: APZ; Writing—original draft preparation: APZ and WHS; Writing—review and editing: APZ, AA, WHS, MA, NRTN, and SJA. NRTN specifically validated the statistical integrity of the study data, preprocessing procedures, and analytical outputs. SJA specifically validated the nutritional interpretation of the study data, preprocessing decisions relevant to nutrition-related variables, and the interpretation of the resulting analytical outputs. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia (Kementerian Pendidikan Tinggi, Sains, dan Teknologi Republik Indonesia) under the BIMA 2026 Penelitian Dosen Pemula scheme, with institutional support from LPPM Universitas Matana.

Data Availability Statement: The respondent-level survey data are not publicly available. Access to the respondent-level dataset is restricted to the research team because it contains information collected from students and their parents or guardians. Results are reported in aggregate form.

Conflicts of Interest: The authors declare no conflicts of interest.

References

[1] C. T. Marx, F. du P. Calmon, and B. Ustun, "Predictive Multiplicity in Classification," in *Proceedings of the 37th International Conference on Machine Learning*, Sep. 2020, vol. 119, pp. 6765–6774. [Online]. Available: https://proceedings.mlr.press/v119/marx20a.html

[2] A. D'Amour *et al.*, "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *J. Mach. Learn. Res.*, vol. 23, no. 226, pp. 1–61, Nov. 2022, [Online]. Available: https://www.jmlr.org/papers/v23/20-1335.html

[3] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J. Clin. Epidemiol.*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/j.jclinepi.2019.02.004.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[5] J. S. Paige *et al.*, "Variability Among Breast Cancer Risk Classification Models When Applied at the Level of the Individual Woman," *J. Gen. Intern. Med.*, vol. 38, Feb. 2023, doi: 10.1007/s11606-023-08043-4.

[6] J. Zhang and T. Li, "Discrimination stability and calibration of cardiovascular risk prediction models in the Framingham baseline cohort," *Sci. Rep.*, vol. 16, no. 1, p. 23965, May 2026, doi: 10.1038/s41598-026-54869-3.

[7] M. Pawelczyk, K. Broelemann, and G. Kasneci, "On Counterfactual Explanations under Predictive Multiplicity," *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, vol. 124. pp. 809–818, Jun. 23, 2020. doi: 10.48550/arXiv.2006.13132.

[8] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *Gigascience*, vol. 6, no. 5, p. gix019, May 2017, doi: 10.1093/gigascience/gix019.

[9] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," *BMC Med.*, vol. 17, no. 1, p. 230, Dec. 2019, doi: 10.1186/s12916-019-1466-7.

[10] T. Takada *et al.*, "Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets," *J. Clin. Epidemiol.*, vol. 137, pp. 83–91, Sep. 2021, doi: 10.1016/j.jclinepi.2021.03.025.

[11] E. W. Steyerberg and F. E. Harrell, "Prediction models need appropriate internal, internal–external, and external validation," *J. Clin. Epidemiol.*, vol. 69, pp. 245–247, Jan. 2016, doi: 10.1016/j.jclinepi.2015.04.005.

[12] R. Saptati, "Pemerintah Salurkan Makan Bergizi Gratis (MBG), Ini Sasaran Utama Penerimanya," *Media Keuangan Kemenkeu*, Feb. 2025. https://mediakeuangan.kemenkeu.go.id/article/show/pemerintah-salurkan-makan-bergizi-gratis-mbg-ini-sasaran-utama-penerimanya
[13] C. Sobek *et al.*, "Purchasing Behavior, Setting, Pricing, Family: Determinants of School Lunch Participation," *Nutrients*, vol. 13, no. 12, p. 4209, Nov. 2021, doi: 10.3390/nu13124209.
[14] L. A. Spruance and T. Vo, "Race/Ethnicity Predicts School Meal Participation," *J. Sch. Health*, vol. 93, no. 4, pp. 305–312, Apr. 2023, doi: 10.1111/josh.13270.
[15] L. A. Spruance, D. Santamaria, and S. Hall, "Parent Perceptions of School Meals as Predictors of Participation Among Pacific Islander Families in the United States," *J. Sch. Health*, vol. 96, no. 6, p. e70169, Jun. 2026, doi: 10.1111/josh.70169.
[16] M. D. Zuercher *et al.*, "Student Perceptions of School Lunch and Their Association With Self-Reported Participation in School Meal Programs," *J. Nutr. Educ. Behav.*, vol. 58, no. 4, pp. 336–345, Apr. 2026, doi: 10.1016/j.jneb.2025.12.012.
[17] S. A. Santana, S. A. Batista, D. da Costa Maynard, V. C. Ginani, R. P. Zandonadi, and R. B. A. Botelho, "Acceptability of School Menus: A Systematic Review of Assessment Methods," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, p. 2242, Jan. 2023, doi: 10.3390/ijerph20032242.
[18] N. S. Guimarães *et al.*, "Plate Food Waste in Food Services: A Systematic Review and Meta-Analysis," *Nutrients*, vol. 16, no. 10, p. 1429, May 2024, doi: 10.3390/nu16101429.
[19] V. T. Siy Van *et al.*, "Predicting undernutrition among elementary schoolchildren in the Philippines using machine learning algorithms," *Nutrition*, vol. 96, p. 111571, Apr. 2022, doi: 10.1016/j.nut.2021.111571.
[20] B. H. Sgura, M. S. M. Saes, C. V. C. Fonseca, and Y. P. B. Adorno, "Heterogeneidade regional no cumprimento das diretrizes do Programa Nacional de Alimentação Escolar: uma análise comparativa entre Nordeste e Sul do Brasil com aprendizado de máquina," *Cad. Saude Publica*, vol. 42, p. e00056125, 2026, doi: 10.1590/0102-311xpt056125.
[21] H. Surendra, A. Sujagat, and G. Wangge, "Mitigating food safety risks in Indonesia's free school meals programme," *Lancet Reg. Heal. - Southeast Asia*, vol. 46, p. 100734, Mar. 2026, doi: 10.1016/j.lansea.2026.100734.
[22] M. A. Wardana and I. Ismail, "Model Explainable Machine Learning untuk Prediksi Kecukupan Gizi Menu MBG," *J. Ilm. Sist. Inf. dan Tek. Inform.*, vol. 9, no. 1, pp. 99–110, Apr. 2026, doi: 10.57093/jisti.v9i1.410.
[23] T. B. Sedayu, N. Saputri, T. P. Akhmaliyah, M. Hanif, and D. Apdillah, "Analisis Prediksi Konsumsi Gizi pada Program Makan Bergizi Gratis (MBG) Menggunakan Metode Machine Learning," *J. Sci. Soc. Res.*, vol. 9, no. 3, pp. 4283–4291, 2026, doi: 10.54314/jssr.v9i3.6644.
[24] I. M. Rachmawati, A. Nurhindarto, and M. T. Sulistyono, "Prediksi Konsumsi Harian Program Makan Bergizi," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 11, no. 1, pp. 2029–2039, Feb. 2026, doi: 10.36341/rabit.v11i1.7609.
[25] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[26] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[27] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction," *Ecol. Modell.*, vol. 411, p. 108815, Nov. 2019, doi: 10.1016/j.ecolmodel.2019.108815.
[28] P. Ploton *et al.*, "Spatial validation reveals poor predictive performance of large-scale ecological mapping models," *Nat. Commun.*, vol. 11, no. 1, p. 4540, Sep. 2020, doi: 10.1038/s41467-020-18321-y.
[29] G. C. Cawley and N. L. C. Talbot, "On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, 2010, [Online]. Available: https://jmlr.org/papers/v11/cawley10a.html
[30] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: https://dl.acm.org/doi/10.5555/3295222.3295230
[31] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.
[32] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems With Shapley-Value-Based Explanations as Feature Importance Measures," in *Proceedings of the 37th International Conference on Machine Learning*, Jun. 2020, vol. 119, pp. 5491–5500. [Online]. Available: https://dl.acm.org/doi/10.5555/3524938.3525447
[33] D. Janzing, L. Minorics, and P. Blobaum, "Feature Relevance Quantification in Explainable AI: A Causal Problem," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020, vol. 108, pp. 2907–2916. [Online]. Available: https://proceedings.mlr.press/v108/janzing20a.html
[34] G. Shmueli, "To Explain or to Predict?," *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, Aug. 2010, doi: 10.1214/10-STS330.
[35] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *SSRN Electron. J.*, 2017, doi: 10.2139/ssrn.3063289.
[36] B. Ustun, A. Spangher, and Y. Liu, "Actionable Recourse in Linear Classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan. 2019, pp. 10–19. doi: 10.1145/3287560.3287566.
[37] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "FACE: Feasible and Actionable Counterfactual Explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb. 2020, pp. 344–350. doi: 10.1145/3375627.3375850.
[38] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki, "The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Aug. 2019, pp. 2801–2807. doi: 10.24963/ijcai.2019/388.
[39] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–29, May 2023, doi: 10.1145/3527848.
[40] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic Recourse," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021, pp. 353–362. doi: 10.1145/3442188.3445899.

[41] N. Hidayat, L. Afuan, and H. R. Jannah, "Prescriptive Learning Analytics for Student Dropout: Integrating Temporal Velocity and Counterfactual Explanations in Longitudinal Data," *J. Comput. Theor. Appl.*, vol. 3, no. 4, pp. 627–646, May 2026, doi: 10.62411/jcta.15920.

[42] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer, "Prediction Policy Problems," *Am. Econ. Rev.*, vol. 105, no. 5, pp. 491–495, May 2015, doi: 10.1257/aer.p20151023.

[43] M. R. Maulana, P. Sudira, P. Priyanto, Y. A. Fadlullah, D. Nurdiana, and A. H. Bin Sofhi @ Subhi, "Explainable Machine Learning for Predicting Indonesian Vocational School Accreditation: Geographic Validation, Probability Calibration, and Subgroup Auditing," *J. Comput. Theor. Appl.*, vol. 4, no. 2, pp. 502–516, Sep. 2026, doi: 10.62411/jcta.17721.

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011, doi: 10.1289/EHP4713.

[45] D. R. Roberts *et al.*, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography (Cop.).*, vol. 40, no. 8, pp. 913–929, 2017, doi: 10.1111/ecog.02881.

[46] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[47] E. W. Steyerberg *et al.*, "Assessing the Performance of Prediction Models," *Epidemiology*, vol. 21, no. 1, pp. 128–138, Jan. 2010, doi: 10.1097/EDE.0b013e3181c30fb2.

[48] M. Pakdaman Naeini, G. Cooper, and M. Hauskrecht, "Obtaining Well Calibrated Probabilities Using Bayesian Binning," *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, Feb. 2015, doi: 10.1609/aaai.v29i1.9602.

[49] C. A. Field and A. H. Welsh, "Bootstrapping Clustered Data," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 69, no. 3, pp. 369–390, Jun. 2007, doi: 10.1111/j.1467-9868.2007.00593.x.

[50] Y. Vergouwe, E. W. Steyerberg, M. J. C. Eijkemans, and J. D. F. Habbema, "Substantial effective sample sizes were required for external validation studies of predictive logistic regression models," *J. Clin. Epidemiol.*, vol. 58, no. 5, pp. 475–483, May 2005, doi: 10.1016/j.jclinepi.2004.06.017.

[51] C. J. Van Rijsbergen, "Evaluation," in *Information Retrieval*, 2nd ed., vol. 30, no. 6, London, UK: Butterworths, 1979, pp. 374–375. [Online]. Available: https://www.dcs.gla.ac.uk/Keith/Chapter.7/Ch.7.html