IPCid

working paper

DOI: http://dx.doi.org/10.38116/ipcid-wp006

IPCid Working Paper **Number 006**
September, 2026

Towards an early-warning system for food insecurity in Brazil: Forecasting vulnerable households using national survey data

Igor Ferrão
socialprotection.org and Institute for Applied Economic Research (Ipea)

Ricardo Campante Vale
Institute for Applied Economic Research (Ipea)

IPCid ipea Institute for Applied Economic Research

IPCid Working Paper No. 006
Towards an early-warning system for food insecurity in Brazil: Forecasting vulnerable households using national survey data
Igor Ferrão and Ricardo Campante Vale

Copyright© 2026
Institute for Applied Economic Research (Ipea)

The International Policy Centre for Inclusive Development (IPCid) is a pillar for the internationalisation of public policy within the Brazilian public sector. The Centre promotes knowledge sharing on public policies between Brazil and its partners, while strengthening South-South cooperation and mutual learning in the Global South.

The text and data in this publication may be reproduced as long as the source is cited.
Reproductions for commercial purposes are forbidden.

The International Policy Centre for Inclusive Development disseminates the findings of its work in progress to encourage the exchange of ideas about development issues. The papers are signed by the authors and should be cited accordingly. The findings, interpretations, and conclusions that they express are those of the authors and not necessarily those of the Centre or the Government of Brazil.

This publication is available online at www.ipea.gov.br. For further information on IPCid publications, please feel free to contact ipcid@ipea.gov.br.

Suggested citation: Ferrão, I., and R. Campante Vale. 2026. "Towards an early-warning system for food insecurity in Brazil: Forecasting vulnerable households using national survey data". Working Paper, No. 6. Brasília: International Policy Centre for Inclusive Development.

DOI: http://dx.doi.org/10.38116/ipcid-wp006

SUMMARY

1 Abstract 5

2 Introduction 5

3 Methodology 8

3.1 Data 8

3.2 Statistical learning approaches 10

4 Results 15

4.1 Logistic regressions 15

4.2 Decision trees 17

4.3 Random forests 17

4.4 Extreme gradient boosting 18

4.5 Support vector machines 19

4.6 LASSO logistic regression 20

5 An assessment of the results 20

5.1 To include or not climatic variables 21

5.2 The trade-offs of model selection 23

5.3 Feature selection or algorithmic flexibility: insights from the LASSO 25

5.4 Threshold choice and policy trade-offs 26

6 Policy discussion 31

7 References 34

Appendix A: hyperparameter tuning 37

LIST OF FIGURES

FIGURE 1. ROC curves (2024 test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2), with each model's CV-optimal threshold *(t*)* marked 27

FIGURE 2. Precision-recall curves (2024 test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2) 28

LIST OF TABLES

TABLE 1. Logistic Regression performance metrics—Original sample (imbalanced 2023 CV) 15

TABLE 2. Logistic Regression performance metrics—Oversampled 2023 CV (*t*: M1 = 0.495, M2 = 0.475, M3 = 0.495, M4 = 0.485*) 16

TABLE 3. Logistic Regression 2024 test performance metrics (*t*: M1 = 0.495, M2 = 0.475, M3 = 0.495, M4 = 0.485*) 16

TABLE 4. Decision Trees: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.385, w/o precip. = 0.385*) 17

TABLE 5. Random Forest: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.205, w/o precip. = 0.225*) 18

TABLE 6. eXtreme Gradient Boosting: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.395, w/o precip. = 0.400*) 19

TABLE 7. Support Vector Machine: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.450, w/o precip. = 0.455*) 19

TABLE 8. LASSO Logistic Regression: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.440, w/o precip. = 0.440*) 20

TABLE 9. Model performance metrics (2024 Test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2) 23

TABLE 10. Threshold sensitivity analysis (2024 Test). Non-precipitation ML variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2) 30

TABLE A1. Hyperparameter tuning: final selected values by model and variant 37

TOWARDS AN EARLY-WARNING SYSTEM FOR FOOD INSECURITY IN BRAZIL: FORECASTING VULNERABLE HOUSEHOLDS USING NATIONAL SURVEY DATA

Igor Ferrão and Ricardo Campante Vale[1]

1 ABSTRACT

Brazil's food insecurity reversed sharply during the COVID-19 pandemic, underscoring the need for effective forecasting tools to protect vulnerable households. Using nationally representative data on food insecurity from Brazil's PNADC (2023–2024), this paper compares logistic regression, machine learning algorithms, and LASSO regression to predict food-insecure households. Machine learning models outperform the traditional logit in prediction: XGBoost identifies about 76 per cent of food-insecure households at 71 per cent accuracy, with the best discrimination of any model (AUROC 0.810). Notably, LASSO regression performs comparably to the non-linear models, indicating that regularised feature selection captures most of the available predictive gains, while algorithmic complexity adds a further, smaller margin of discriminative power. Rainfall data, contrary to expectations, added no predictive value. These findings inform the potential design of a national system to support food security policy targeting.

Keywords: food insecurity, predictive modelling, policy targeting, early-warning systems, Brazil

2 INTRODUCTION

The COVID-19 pandemic demonstrated that the fight against hunger is not linear, remaining one of the most pressing global policy challenges of our time. As highlighted by the United Nations (2023), compounding exogenous shocks, such as the pandemic, armed conflicts, climate change, and widening inequalities, pushed global hunger in 2022 back to levels not seen since 2005. Despite modest signs of improvement between 2022 and 2024, an estimated 638 to 720 million people still faced hunger in 2024, remaining persistently above 2015 levels, while nearly 2.3 billion people experienced moderate or severe food insecurity (United Nations 2025). These figures underscore the immense challenges that remain in achieving Sustainable Development Goal 2 (SDG 2), "Zero Hunger," by 2030.

1. We are grateful to Douglas Braga (Ipea), Lourenço Araújo (UFMG), Carlo Cafiero (FAO), and Hélio Macêdo Filho (MDS), whose comments and feedback greatly helped to improve this study. All remaining errors are our own. The findings, interpretations, and conclusions expressed in this paper are strictly those of the authors and do not necessarily reflect the official views or position of the International Policy Centre for Inclusive Development (IPCid), the Institute for Applied Economic Research (Ipea), or the Government of Brazil.

In Brazil, after successfully exiting the UN's Hunger Map in 2014, the country experienced a drastic reversal during the pandemic. This regression was driven by the socioeconomic fallout of the COVID-19 pandemic, macroeconomic volatility, and shifts in public expenditures allocated to food and nutrition security since 2016, transforming hunger into an acute national crisis (Rede PENSSAN 2022; Ribeiro-Silva et al. 2020). For instance, between the 2014–2016 and 2020–2022 periods, the prevalence of severe food insecurity surged from 1.9 to 7.3 per cent, while moderate or severe rose from 18.3 to 28.9 per cent (FAO et al. 2022). In headcount terms, this amounts to roughly 20 million more people in moderate or severe food insecurity over half a decade. By 2024, national indicators of undernourishment and food insecurity returned to levels comparable to those recorded in 2014 (FAO et al. 2025).

These trajectories illustrate that national food security is highly vulnerable to sudden, exogenous shocks. Whether triggered by a global health crisis like the COVID-19 pandemic, by intensifying climate-related disruptions or by geopolitical conflicts, these events can rapidly destabilise vulnerable populations. Worldwide, extreme weather events such as acute rainfall and subsequent flooding, or severe agricultural and ecological droughts are increasing in frequency, severity, and duration (UNEP 2025). At the time of writing (2026), the war in the Middle East is pushing fuel prices up, threatening a new rise in food prices. Consequently, developing means of forecasting food crises and those at risk remain as essential as ever to preserve the right to adequate food and decent lives.

Predicting which households are food-insecure can enable rapid and effective policy targeting in response to shocks. Traditionally, the standard analytical approach for this in the specialised literature has been the logistic regression (LR) model. However, most existing analyses that use this method aim to study the determinants of food insecurity, rather than optimising for predictive power (Bakhtsiyarava et al. 2021). As the policy focus shifts from simply understanding past drivers to actively forecasting future crises, recent literature has demonstrated that machine learning (ML) techniques can yield exceptionally strong predictive capabilities (Kleinberg et al. 2015; Mullainathan and Spiess 2017).

For example, studies have shown that ML algorithms perform effectively when predicting both child and adult malnutrition, as measured by standard performance metrics such as accuracy, precision, and recall (sensitivity) (Janssen et al. 2024; Rao et al. 2025).[2] Saihood and Sonuç (2022) compared the performance of eight different ML algorithms to predict anaemia in children using social factors as features. The authors observed that these techniques present good results for this prediction problem, especially the decision tree (DT), random forest (RF) and multilayer perceptron algorithms.

Historically, a significant barrier to applying predictive models to forecast food insecurity was the lack of high-quality, nationwide, periodically collected data to form a time series. To fill this gap, the Department of Evaluation, Information Management, and the Single Registry (SAGICAD)[3] from the Ministry of Social Development (MDS) established a partnership with the Brazilian Institute of Geography and Statistics (IBGE) in 2023. Together, the two institutions introduced the Brazilian Scale of Food Insecurity (EBIA) into Brazil's Continuous National

2. Features within a ML context are the input variables used in the models. They are also known as predictors or independent variables (James et al., 2021).

3. *Secretaria de Avaliação, Gestão da Informação e Cadastro Único* in Portuguese.

Household Sample Survey (PNADC). The EBIA was introduced in the questionnaires applied during the fourth quarter of 2023 and 2024.

This is the first time that nationally representative panel data on food insecurity have been available at this scale.[4] The PNADC, in its most recent design, included a module to measure food insecurity, marking the return of EBIA to a nationwide household survey after five years without such data. Unlike previous national surveys using the EBIA (such as PNAD and POF[5]) that only allowed for estimates at highly aggregated spatial levels, PNADC statistics can be estimated at the geographic stratum level, enabling us to link food insecurity with local climate data. The availability of this new data provides a crucial opportunity to evaluate the performance of different predictive models when trying to predict food insecurity on the entire Brazilian population and territory.

This study leverages this new, nationally representative data to formally evaluate and construct predictive models that can subsidise Brazil's CadINSAN indicator, which aims to forecast the number of food-insecure families at the municipal level (MDS 2025). In doing so, this paper contributes to the applied public policy literature by providing a large-scale, out-of-sample framework that can contribute to the future endeavour of building an effective food insecurity early-warning system in Brazil.

To the best of our knowledge, this is the first study to apply a suite of complex statistical learning algorithms to predict food insecurity in Brazil at a national scale. While Barbosa and Nelson (2016) used a support vector machine (SVM) to assess food security in Ceará, they restrict their scope to one model and one state. Previous literature widely applied logit or standard linear regression with the purpose of identifying determinant factors rather than making predictions (Campante Vale 2026; Cherol et al. 2022; Hoffmann 2008, 2013, 2021; Pontes et al. 2018).

We systematically explore how different ML algorithms perform across standard predictive metrics when fed with different model specifications and temporal realities, and benchmark them against the LR. To explore our predictive power, we act as if we were to forecast food-insecure households in 2024. First, we train our model using the 2023 data, registering its performance through a cross-validation approach. Then, we test the predictions on the 2024 data.

The results of this analysis revealed an operational trade-off between policy reach and targeting efficiency. While the traditional LR and certain ensemble ML algorithms struggled to balance these two competing needs, XGBoost emerged as a robust, balanced alternative. It successfully identifies 76.3 per cent of all truly food-insecure households while maintaining 71.2 per cent overall accuracy and achieving the highest discriminative power of any model evaluated. The RF follows closely in overall discrimination, trailing only XGBoost. Crucially, the LASSO yielded performance comparable to non-linear methods, providing a highly viable option that keeps feature selection both tractable and methodologically stable.

Furthermore, this framework allows us to formally test if we can improve our predictive power by integrating rainfall data, as suggested by earlier work (Campante Vale 2026). Contrary

4. To be sure, in PNADC rotation scheme, one-fifth of the sample is replaced at each quarter, after completing five interviews. Thus, between 2023Q4 and 2024Q4, roughly twenty per cent of the households should be the same (in their 1st and 5th interviews), while the other eighty per cent are actually pooled cross-sections.

5. PNAD is the earlier version of the Brazilian National Household Survey *(Pesquisa Nacional de Amostra por Domicílios)*, which was not conducted on a rolling basis as its current "Continuous" version. POF is Household Budget Survey of Brazil *(Pesquisa de Orçamento Familiar)*.

to initial expectations, we found that integrating rainfall data offered no significant predictive enhancement. During the internal 2023 cross-validation, the algorithms' performances remained practically equivalent with or without these variables. When deployed on the unseen 2024 data, these metrics actively destabilised certain models, triggering more misclassifications when confronted with a new climatic reality.

Besides this current section, the immediately following section details the data sources and methodology, including an explanation of the selected statistical approaches, the performance metrics used, the cross-validation technique applied, as well as the test with the 2024 data. Lastly, we present the performance metrics obtained for each of our models and discuss the implications of these results for policymaking.

3 METHODOLOGY

This section describes the data and statistical learning framework used to perform the analysis. The data subsection introduces the two primary sources, Brazil's Continuous National Household Sample Survey (*PNAD Contínua*) and the CHIRPS satellite-derived precipitation data. The second subsection then details the supervised learning algorithms considered, spanning logistic regression and a regularised variant, non-parametric tree-based ensembles, and a support vector machine. This subsection also explains the common evaluation criteria, the models' calibration procedures, and the temporal hold-out testing strategy applied uniformly across all specifications.

3.1 DATA

3.1.1 PNAD Contínua

Our analysis is built on the official data from Brazil's Continuous National Household Sample Survey (PNADC), collected during the fourth quarter of 2023 and 2024. The PNADC is the country's primary tool for tracking socioeconomic and demographic trends, employing a sophisticated multistage sampling design that allows for a deep and evolving understanding of the population. Its design, where households are interviewed quarterly over five quarters, not only ensures our results are statistically representative of the entire nation but also allows for the inclusion of timely, policy-relevant supplementary research modules.

It was through this modular structure that the Brazilian Institute of Geography and Statistics (IBGE) introduced the official Brazilian Scale of Food Insecurity (EBIA) into the survey. This decision was made in response to growing evidence suggesting a rise in food insecurity following the COVID-19 pandemic. The inclusion of the EBIA in late 2023 represented a landmark return of this official tool to nationwide IBGE surveys after five years, renewing Brazil's capacity to generate timely and authoritative evidence on food security.

The EBIA itself is a validated questionnaire that assesses a household's real-world experiences with food access and affordability over the preceding three months (Reichenheim et al. 2016). Based on the number of affirmative answers, households are classified into one of four categories: food security, mild food insecurity, moderate food insecurity, or severe food insecurity. The questionnaire is adapted for different household types, using a more detailed 14-item version for those with children and a shorter 8-item version for adult-only households (Segall-Corrêa et al. 2014).

The purpose of this predictive analysis is to develop a model capable of reliably identifying households unequivocally facing nutritional hardship. To this end, we combine the moderate and severe food insecurity categories into a single binary outcome, representing households that experience significant food constraints. This approach simplifies interpretation while maintaining a clear focus on the most at-risk groups.

Aiming to attain operational viability for a potential policy application in the future, we deliberately restricted our predictor set strictly to variables compatible with *Cadastro Único* (CadÚnico)—Brazil's national social registry—even though the PNADC offers a much broader array of household characteristics. While PNADC is a survey that ensures respondents are never identified, even for policymakers, CadÚnico is a tool that actively informs policy targeting.

This approach mirrors the methodology of CadINSAN, the municipal-level indicator that initially motivated this research. CadINSAN trains a logistic regression on the PNADC to identify predictive factors and then applies that model to *Cadastro Único* using only the variables the two databases share (MDS 2025). By matching our feature set with the social registry from the start, we guarantee that our results remain fully deployable for this type of real-world policy application.

3.1.2 CHIRPS

In addition to well-established socioeconomic drivers, a key question for this analysis is whether recent climate conditions add significant predictive power to our models of food insecurity. In a country as large as Brazil, local rainfall patterns can heavily influence agricultural output, which in turn affects food supply and prices. Therefore, we tested the hypothesis that incorporating recent rainfall data could improve our ability to identify at-risk households.

To measure this potential impact, we used data from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS). This is a globally recognised, high-resolution dataset developed by UC Santa Barbara and the U.S. Geological Survey, with a record spanning from 1981 to December 2024. The strength of CHIRPS lies in its hybrid methodology, which provides a more complete and reliable picture of rainfall than either satellite data or ground stations could alone. It works by blending satellite imagery that estimates rainfall by observing the temperature of cloud tops with data from thousands of *in situ* weather stations across the globe, which corrects biases in the satellite data and improves overall accuracy.[6]

Using the Google Earth Engine platform, we recover the data and create an indicator of cumulative rainfall during the relevant period for agricultural droughts in Brazil. Following Campante Vale (2026), we calculated the cumulative average rainfall for each of the PNADC's geographic survey areas during the period of November 2022 to April 2023. This timeframe was deliberately chosen as it corresponds to the main agricultural season in Brazil, the success of which directly influences the food supply and market prices that households would have experienced in the months leading up to the food security survey in late 2023. Nonetheless, Campante Vale (2026) suggests that varying the time span over the year preceding the survey to determine the relevant rainfall measurement does not change the results.

6. More information on CHIRPS available at: https://www.chc.ucsb.edu/data/chirps3.

3.2 STATISTICAL LEARNING APPROACHES

3.2.1 Model specifications

The first statistical learning approach used to predict food-insecure households was the logistic regression (LR), a well-established methodology in the literature to make inferences and predictions about food insecurity.

To ensure the robustness of our findings, we employ a forward selection approach. We begin with a baseline specification and then systematically introduce additional sets of control variables, as established in the literature. This method allows us to verify that our core results remain consistent across different model specifications and to account for potential omitted variables bias. The logit can be expressed by the following expression:

$$y_{l,d} = \alpha + Z_{l,d}\theta_{rainfall} + X_{1l,d}\beta_1 + X_{2l,d}\beta_2 + X_{3l,d}\beta 3 + \varepsilon_{l,d} \quad (1)$$

The $y_{l,d}$ is the log-odds (logit) of being food insecure.[7] Then, our baseline specification employs only the Z variables, which represent the rainfall levels experienced by a household d residing in a geographical stratum l. We test for both linear and quadratic forms to allow for non-linear relationships, an approach that is not only generally supported by the literature but also confirmed by our data. The vector θ contains the parameters of primary interest, which quantify the association between rainfalls and the probability of a household being moderately or severely food-insecure.

Our next specification introduces a well-established set of socioeconomic controls (X_1): household income per capita (logarithmic form); the race, gender and educational level of the household head;[8] and whether the household is located within an urban or rural area. The subsequent specification incorporates occupational control variables (X_2), comprised of dummy variables indicating the occupational status of the household head.[9] Our final and most comprehensive specification includes regional variables (X_3) to account for heterogeneity across Brazil's vast and diverse regions. Finally, β_i, $i = 1, 2, 3$, is the vector of parameters associated with the control variables X_i, while ε is an idiosyncratic error term that is assumed to follow a logistic distribution. We can summarise our models as the following:

1. $M1: y_{l,d} = \alpha + Z_{l,d}\theta_{rainfall} + X_{1l,d}\beta_1 + \varepsilon_{l,d}$, with both linear and quadratic forms of $Z_{l,d}$.
2. $M2: y_{l,d} = \alpha + Z_{l,d}\theta_{rainfall} + X_{1l,d}\beta_1 + X_{2l,d}\beta_2 + \varepsilon_{l,d}$, with both linear and quadratic forms of $Z_{l,d}$.
3. $M3: y_{l,d} = \alpha + Z_{l,d}\theta_{rainfall} + X_{1l,d}\beta_1 + X_{2l,d}\beta_2 + X_{3l,d}\beta_3 + \varepsilon_{l,d}$, with both linear and quadratic forms of $Z_{l,d}$.
4. $M4: y_{l,d} = \alpha + X_{1l,d}\beta_1 + X_{2l,d}\beta_2 + X_{3l,d}\beta_3 + \varepsilon_{l,d}$, without precipitation as predictor.

7. Formally, it is $\ln\left(\frac{p}{1-p}\right)$, in which p is the probability of being food insecure.

8. Incomplete elementary, complete elementary, incomplete high school, complete high school, incomplete higher education, and complete higher education.

9. Three different occupational dummies are considered: occupied, formal occupation, and retired.

As described above, the intent of this study is to evaluate the performance of alternative statistical learning approaches when compared to the traditional methodology employed by the literature, the LR. Such algorithms are fundamentally different in their philosophy when compared to standard econometric models. For instance, the LR is a parametric model that assumes a specific functional form for f. As its name implies, it uses the logistic function to ensure that the estimated probability of Y belonging to a particular class, given X, always falls between 0 and 1. In contrast, our selected ML approaches are non-parametric, meaning they do not impose explicit assumptions about the functional form of f.[10] This gives them considerable flexibility, enabling them to learn complex, non-linear patterns directly from the data. Instead of providing a fixed specification, we provide them with the full set of potential predictors, and they identify the most important predictive patterns directly on their own.

Supported by recent studies on malnutrition prediction (Janssen et al. 2024; Rao et al. 2025), we selected four algorithms that show promising results in the literature: Decision Tree (DT), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). Finally, to assess whether these models outperform the LR due to their access to the full feature set or due to non-linear flexibility, we also estimate a Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression.

3.2.2 Performance metrics

To compare the models' performances in our analysis, a confusion matrix was constructed for each one of them. A confusion matrix is a table that evaluates a model's performance by comparing its predictions to the actual outcomes. It is comprised of four different components: i) True Positives (TP) are food-insecure households correctly predicted by the model; ii) True Negatives (TN) are food-secure households the model correctly identified; iii) False Positives (FP, type I error), which occur when the model incorrectly flags a food-secure household as at-risk; and iv) False Negatives (FN, type II error), which happen when the model incorrectly flags a food-insecure household as secure.

From these components, we can calculate key performance statistics like accuracy, which represents the total percentage of correct predictions by the model for all households.[11] However, as highlighted by James et al. (2021), a model can have a high accuracy and still perform poorly. For example, a model could achieve a high accuracy on our imbalanced data by simply always predicting the majority food-secure class. Such a model would appear successful while failing to find anyone in need. To avoid misleading conclusions, accuracy should be complemented with additional performance statistics that move beyond a simple "right vs. wrong" score.

This is why the model's predictive mistakes (FP and FN) are so important. For a policy designed to deliver aid, these two types of errors are not equal. It is preferable to incorrectly flag a secure household (FP) than it is to miss a household that is in need (FN), since the social cost of a missed family is far greater than the operational and financial costs of a FP. As a result, we must prioritise a metric that directly measures the model's success in finding the

10. The link between X and Y can be described by a function f, which represents the predictable part of Y that can be explained by X. Any remaining variation is represented by an error term ε. As such, the relationship between X and Y can be written as the general form: $Y = f(X) + \varepsilon$.

11. $Accuracy = (TP + TN)/(TP + TN + FP + FN)$.

target at-risk population. This primary metric is sensitivity, also known as recall or true positive rate, which represents the proportion of all truly food-insecure households that the model successfully identifies.[12]

Alongside sensitivity and accuracy, we also report on precision, which denotes the positive predictive value of the model, i.e., the proportion of positive predictions that were correct. It is obtained by dividing the number of correctly identified food-insecure households by the total number of households the model predicted as food-insecure.[13] In simpler terms, precision shows us how trustworthy the model is when it flags a household as at-risk.

3.2.3 Class imbalance and hyperparameter tuning

While flexible, the predictive power of these algorithms faces a significant challenge when the class of interest is under-represented. The PNADC data contains a relatively small proportion of food-insecure households—approximately 10 per cent of raw observations, falling to 9.44 per cent when survey weights are applied. This imbalance is a critical problem for any predictive model, as it hinders the algorithms' ability to learn the subtle patterns identifying the vulnerable households we are most interested in. If left unaddressed, models simply learn to always predict the majority class (food-secure) to artificially inflate their overall accuracy, whilst failing to identify almost anyone actually in need.

To resolve this, we employed oversampling to the training data (2023), a standard statistical technique that creates a balanced training dataset by randomly replicating the rarer food-insecure observations until they match the frequency of the secure group. By replicating the rarer food-insecure observations, the new balanced environment allows the models to properly learn the specific socioeconomic and demographic patterns that distinguish a vulnerable household from a secure one, rather than simply ignoring these subtle signals to minimise overall error rates.

While we acknowledge that our oversampling relies on row duplication, which can increase the susceptibility of complex, tree-based algorithms to overfitting, it was deliberately chosen over synthetic generation methods like the Synthetic Minority Over-sampling Technique (SMOTE). The latter implicitly assumes a continuous feature space and generates new observations via interpolation between nearest neighbours. Given that our dataset is composed of many categorical and binary socioeconomic variables (e.g., race, formal employment status, and region), such interpolation can yield implausible or internally inconsistent demographic profiles, thereby violating the underlying data-generating process. We also tested class weights, applied per algorithm, as an alternative to oversampling, but it has not improved our prediction objectives.[14]

Finally, another key question when applying complex machine learning algorithms (DT, RF, XGBoost, and SVM) is the hyperparameter tuning. Unlike standard parameters that a model learns directly from the data during training, hyperparameters are overarching configurations

12. *Sensitivity = TP/(TP + FN).*

13. *Precision = TP/(TP + FP).*

14. The full inverse-class-weight ratio pushed the Decision Tree and Random Forest toward predicting the positive class almost everywhere; a square-root-dampened weight corrected this for the Random Forest but not the Decision Tree and left the remaining algorithms' metrics unchanged or worse than under oversampling. We kept oversampling as the primary approach for all six models.

established by the researcher beforehand. They dictate the fundamental behaviour and architecture of the algorithm, controlling critical aspects such as its structural complexity, its learning rate, and its safeguards against overfitting. Identifying the optimal combination of these settings is important, as inappropriate hyperparameters can severely compromise an algorithm's predictive power. Appendix A details the tuning strategies applied to each algorithm and reports the final selected hyperparameters.

3.2.4 Training and testing strategies

To ensure the models' effectiveness and reliability, we implemented a two-stage evaluation strategy: an internal validation using the 2023 data, followed by an external validation against the unseen 2024 dataset. This can be seen as a forecasting task simulation. First, the models are trained using the data available in 2023 and their performances are assessed in the same year. Then, we predict household food insecurity using 2024 data, as if we could not see the true response. Finally, we assess the temporal generalisability of our findings, explicitly testing whether the algorithm that performed best during the 2023 CV maintains its predictive lead on the unseen 2024 test set.

For the internal validation of the models, we applied a 5-fold Cross-Validation (CV) approach on the 2023 data. The data was partitioned into five equal subsets: iteratively, four folds (80 per cent of data) were used to train the models while the remaining fold (20 per cent) was held out purely for testing.[15] Crucially, to preserve the integrity of the evaluation, oversampling was applied exclusively to the training data within each specific iteration. The 20 per cent held-out test fold remained completely untouched in its original, imbalanced state. This ensures that while the algorithms learn from a balanced environment, their performance metrics reflect true, real-world conditions. This process is repeated five times, so that every household is held out and predicted on exactly once. Performance metrics are then computed a single time on this complete set of out-of-fold predictions, giving a reliable estimate of model stability that is not inflated by any single random data split.

Following the internal CV, we proceeded to the external validation to verify the real predictive power of our models. To construct the final "champion" models, we applied a second oversampling procedure to the entirety of the 2023 data, creating a fully balanced set comprising all available observations from that year. The algorithms were trained on this full sample to improve their learning potential. Subsequently, these final selected models were tasked with predicting food insecurity in the entirely separate, imbalanced 2024 PNADC dataset, providing a temporally unbiased evaluation of the models' capacity to generalise to new populations. As is standard in supervised learning, these predicted classifications were then compared against the households' actual observed status to assess model performance.

All tests' performance metrics were calculated using the official PNADC household survey weights on the original, imbalanced test distributions. This ensures that our reported results, whether derived from the pooled out-of-fold predictions of the 2023 CV or from the final 2024 external validation, remain strictly representative of the true demographic reality of the Brazilian population, offering a reliable and scalable tool for policymaking at the national level.

15. To ensure strict comparability, the 2023 folds (including all tuning subsamples) were partitioned prior to training. Every model was, then, trained and evaluated using these exact same partitions. Any performance differences that arise, then, may be attributed to the algorithms themselves, not from how the data happened to be split.

3.2.5 Calibrating the classification thresholds

Thus far, we have referred to model outputs as if they strictly assign a positive or negative binary class to each observation. In reality, models generally estimate the continuous probability of belonging to one class over the other. A straightforward way to convert these probabilities into predicted classes is by majority rule; in our binary setting, this means classifying as food insecure those with predicted probabilities above 50 per cent. However, this naive threshold does not necessarily optimise performance metrics in our context of predictions. Therefore, rather than assuming a fixed cutoff, we empirically calibrate the probability threshold to optimise these metrics.

Our strategy was to determine each algorithm's optimal operating threshold (t^*) during CV by maximising Youden's J ($J = sensitivity + specificity - 1$)[16] on the model's CV receiver operating characteristic (ROC) curve (Youden 1950). This approach yields a mathematically defined threshold rather than one based on an arbitrary, hand-picked target for either metric. As the food-insecurity prevalence in our sample is low, maximising J implicitly weights FN more heavily than FP (Smits 2010). This dynamic represents the ideal scenario for the early-warning tool we aim to develop, as the human cost of missing a genuinely vulnerable household far outweighs the administrative cost of flagging a secure one for potential intervention. This single, calibrated value then carries forward unchanged into the 2024 external test, serving as the decision rule for classifying households in both the CV phase and the 2024 external test, including the SVM's Platt-scaled probabilities.[17]

However, because each algorithm's t^* is calibrated to its own specific probability distribution, comparing models exclusively at their respective optima can obscure how each one trades sensitivity for precision away from that point. To address this, we also evaluate every model's performance across a standardised grid of fixed classification thresholds (ranging from 0.20 to 0.50). This allows us to compare the six algorithms' behaviour side by side at identical cut-offs, rather than only at their individual calibrated thresholds.

Finally, we also report the area under the ROC curve (AUROC), a measure of the model's discriminative power that complements the threshold-dependent metrics above. The ROC curve traces sensitivity against the FP rate across every possible classification threshold, and the area under it condenses that trade-off into a single, threshold-independent number. This number has an intuitive interpretation: the probability that the model assigns a higher risk score to a randomly chosen food-insecure household than to a randomly chosen food-secure one. A value of 0.5 corresponds to random guessing (the model would rank such a pair correctly only half the time), while a value of 1 indicates perfect separation between the two groups. As such, a higher AUROC means the model separates food-insecure from food-secure households more cleanly at every possible threshold, providing greater flexibility to raise sensitivity without triggering a matching rise in false positives.

16. Specificity is the true negative rate: $TN/(TN + FP)$.

17. Because the SVM algorithm natively generates a decision boundary rather than probability scores, we apply Platt scaling to its fitted values (Platt, 1999). This procedure fits a logistic function that yields probability estimates compatible with the same t^*-based classification framework used for the other five models.

4 RESULTS

This section reports the predictive performance for the selected algorithms, in order: LR, DT, RF, XGBoost, SVM, and LASSO logistic regression. Each subsection presents results for both precipitation and non-precipitation variants, estimated on the 2023 CV training sample and evaluated on the 2024 external test set.

4.1 LOGISTIC REGRESSIONS

Our initial results, presented in Table 1, clearly demonstrate why applying predictive models to highly imbalanced data fails to produce useful tools for policy targeting. Across all baseline LR specifications evaluated via CV on the original 2023 sample, the algorithms were extremely conservative when predicting a household as food-insecure, something that can be inferred by the extremely low FP rates of virtually 0 per cent for all specifications. The algorithms are anchored to optimise overall accuracy and, consequently, apply the easiest way to achieve it in this imbalanced dataset, which is predicting the majority class (food-secure) for almost every observation.

TABLE 1. Logistic Regression performance metrics—Original sample (imbalanced 2023 CV)

Metric	(M1) Rainfall + demographic Controls	(M2) M1+ labour Controls	(M3) M2 + region Controls	(M4) All controls but rainfall
Sensitivity	0.0185	0.0155	0.0142	0.0134
Accuracy	0.8982	0.9001	0.9018	0.9018
Precision	0.1611	0.1748	0.2066	0.1997
False Negative Rate	0.9815	0.9845	0.9858	0.9866
False Positive Rate	0.0100	0.0077	0.0057	0.0056

Notes: This table presents performance metrics of a logistic regression on predicting food-insecure households in PNADC, assessed via cross-validation (2023) on the original imbalanced sample. Metrics weighted by household survey weight. Abbreviations: CV, cross-validation.

Therefore, while the logistic specifications boast impressive accuracies of approximately 90 per cent, the sensitivities of less than 2 per cent for all of them represent a major issue for the aim of our analysis. This means that from all truly food-insecure households, the models identified virtually none, rendering them functionally unfeasible for targeting within a policy context. This outcome explicitly illustrates the necessity of our oversampling methodology to properly train the algorithms.

The results for the oversampled data, shown in Table 2, confirm that using a balanced dataset drastically improved the models' ability to identify food-insecure households. Sensitivity skyrockets from near zero to values ranging between 73 per cent and 76 per cent. This means the models are now capable of correctly identifying a substantial majority of the truly food-insecure households.

This improvement in identifying vulnerable households is driven by a drastic reduction in the FN rate across all specifications. However, this adjustment also results in a concurrent increase in the FP rate as a consequence of the wider consideration that the algorithm is making, meaning the models now incorrectly flag a higher proportion of food-secure households as being at risk.

TABLE 2. Logistic Regression performance metrics—Oversampled 2023 CV (*t*: M1 = 0.495, M2 = 0.475, M3 = 0.495, M4 = 0.485*)

Metric	(M1) Rainfall + demographic Controls	(M2) M1+ labour Controls	(M3) M2 + region Controls	(M4) All controls but rainfall
Sensitivity	0.7389	0.7638	0.7285	0.7374
Accuracy	0.6733	0.6527	0.6874	0.6779
Precision	0.1877	0.1817	0.1934	0.1898
False Negative Rate	0.2611	0.2362	0.2715	0.2626
False Positive Rate	0.3335	0.3588	0.3168	0.3282
AUROC	0.7699	0.7703	0.7723	0.7709

Notes: This table presents performance metrics of a logistic regression predicting food-insecure households in PNADC, assessed via cross-validation (2023) on the oversampled folds. Metrics weighted by household survey weight. Abbreviations: CV, cross-validation.

While the CV phase demonstrates that the oversampling methodology improves the algorithm's learning process, the true measure of a model's utility lies in its performance on completely unseen, future data. Table 3 shows that the four logistic specifications generalise well from the 2023 CV training to the 2024 test. Sensitivity holds nearly steady for M1 (73.9 per cent to 73.5 per cent) and for M2, the best performing specification in terms of sensitivity (76.4 per cent to 76.3 per cent).

However, this stability was not shared evenly across the four specifications. The M4 specification exhibits a more pronounced decrease, with sensitivity declining by 2.5 p.p. (from 73.7 per cent to 71.2 per cent). Finally, M3 shows the weakest generalisation of the four, falling by 4.9 p.p. (from 72.9 per cent to 67.9 per cent). Precision tells a more consistent story between years, as all four specifications lose roughly three to four percentage points between the internal CV and the 2024 test, regardless of the specific controls included.

TABLE 3. Logistic Regression 2024 test performance metrics (*t*: M1 = 0.495, M2 = 0.475, M3 = 0.495, M4 = 0.485*)

Metric	(M1) Rainfall + demographic Controls	(M2) M1+ labour Controls	(M3) M2 + region Controls	(M4) All controls but rainfall
Sensitivity	0.7347	0.7631	0.6792	0.7124
Accuracy	0.6567	0.6375	0.7048	0.6902
Precision	0.1471	0.1439	0.1599	0.1581
False Negative Rate	0.2653	0.2369	0.3208	0.2876
False Positive Rate	0.3497	0.3728	0.2930	0.3115
AUROC	0.7540	0.7581	0.7594	0.7663

Notes: This table presents performance metrics of a logistic regression predicting food-insecure households in PNADC, assessed via external validation on the unseen data (2024). Metrics weighted by household survey weight; AUROC, Area Under the Receiver Operating Characteristic curve.

Even though M2 possesses the highest sensitivity across all specifications, it is M4 that shows the highest AUROC at 0.766. Independent of its specific optimal threshold ($t^* = 0.485$), this AUROC means that if we randomly select one genuinely food-insecure household and one food-secure one, M4 will correctly assign a higher risk probability to the vulnerable household roughly 76.6 per cent of the time.

4.2 DECISION TREES

Following the standard approach of the LR, we turn to the DT algorithm to evaluate whether adopting a non-parametric ML approach can enhance the predictive power, even if we are using a relatively simple ML model. Table 4 presents the DT performance metrics for both the internal 5-fold CV and the external 2024 validation test, comparing variants with and without precipitation variables.

The results reveal that the DT algorithm represents a compelling alternative to the LR due to its reliability and strong performance. While the evaluation on completely unseen data revealed minor fluctuations, the metrics remained relatively stable across temporal contexts. Specifically, sensitivity declined from about 78.5 per cent in the internal CV for both specifications to 73.8 per cent and 73.5 per cent in the 2024 test for the non-precipitation and precipitation-inclusive variants, respectively, while accuracy increased approximately 3 p.p. for both.

TABLE 4. Decision Trees: CV (2023) and final evaluation (2024) performance metrics
(*t*: w/precip. = 0.385, w/o precip. = 0.385*)

Metric	DT w/ rainfall (CV)	DT w/o rainfall (CV)	DT w/ rainfall (2024 Test)	DT w/o rainfall (2024 Test)
Sensitivity	0.7840	0.7846	0.7351	0.7375
Accuracy	0.6471	0.6456	0.6758	0.6749
Precision	0.1821	0.1815	0.1550	0.1550
False Negative Rate	0.2160	0.2154	0.2649	0.2625
False Positive Rate	0.3672	0.3689	0.3291	0.3302
AUROC	0.7696	0.7687	0.7596	0.7596

Notes: This table presents performance metrics of a decision tree algorithm predicting food-insecure households in PNADC. Columns refer to models with and without precipitation variables assessed via internal cross-validation (2023) and external validation on the unseen 2024 data. Metrics weighted by household survey weight. Abbreviations: DT, Decision Tree; CV, cross-validation; w/, with; w/o, without; precip., precipitation; AUROC, Area Under the Receiver Operating Characteristic curve.

Notably, the optimal decision threshold for both DT variants ($t^* = 0.385$) requires a lower cut-off than the LR specifications. Finally, the DT demonstrates a solid overall discriminatory power, yielding an AUROC of 0.760 and performing on par with the other algorithms.

4.3 RANDOM FORESTS

To build upon the single DT, we implemented an RF algorithm. As a natural mathematical expansion of the DT, the RF creates an ensemble of numerous independent DTs and aggregates their predictions to reduce overall model variance (James et al. 2021). Table 5 presents the RF performance metrics across the internal CV and the external 2024 test, comparing specifications with and without precipitation predictors.

At its own optimal threshold, notably lower than those of the LR and DT ($t^* = 0.205$ with precipitation, $t = 0.225$ without), the RF achieves a high 2024 sensitivity of 72.5 per cent and 74.5 per cent, respectively, within the same range as the other algorithms. The RF also posts

some of the highest accuracies among all algorithms for both specifications (72.2 per cent without rainfall predictors and 73.4 per cent with). Both variants also show an exceptionally high AUROC of approximately 0.794, providing evidence of robust discriminatory power.

TABLE 5. Random Forest: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.205, w/o precip. = 0.225*)

Metric	RF w/ Precip. (CV)	RF w/o Precip. (CV)	RF w/ Precip. (2024 Test)	RF w/o Precip. (2024 Test)
Sensitivity	0.7951	0.7792	0.7250	0.7450
Accuracy	0.6812	0.6903	0.7344	0.7219
Precision	0.2004	0.2030	0.1835	0.1793
False Negative Rate	0.2049	0.2208	0.2750	0.2550
False Positive Rate	0.3307	0.3189	0.2649	0.2800
AUROC	0.7962	0.7933	0.7936	0.7938

Notes: This table presents performance metrics of a random forest algorithm predicting food-insecure households in PNADC. Columns refer to models with and without precipitation variables assessed via internal cross-validation (2023) and external validation on the unseen 2024 data. Metrics weighted by household survey weight. Abbreviations: RF, Random Forest; CV, cross-validation; w/, with; w/o, without; precip., precipitation; AUROC, Area Under the Receiver Operating Characteristic curve.

Similarly to the DT, the RF algorithm shows some performance variation between the 2023 training and the 2024 test. The specification without precipitation shows a modest decline in sensitivity across time, falling by 3.4 p.p., while the precipitation-inclusive variant declines more, by 7.0 p.p.

4.4 EXTREME GRADIENT BOOSTING

Continuing with tree-based ensembles, we now turn to XGBoost, which combines many DT through a different mechanism than the RF. Rather than averaging independent trees, XGBoost grows them sequentially, with each new tree designed to correct the errors of the ensemble built so far, ultimately combining their predictions (Chen and Guestrin 2016).

As shown in Table 6, the XGBoost algorithm demonstrates a very strong overall performance. It reaches a sensitivity of 76.3 per cent in 2024 for the specification without precipitation, matching the best performing LR specification (M2). It also maintains high accuracy at its optimal thresholds: 75 per cent with rainfall predictors ($t^* = 0.395$) and 71 per cent without them ($t^* = 0.400$). The former represents the highest accuracy recorded across all models at their respective optimal cut-offs. Additionally, the model also delivers strong precision (19.2 per cent with precipitation and 17.6 per cent without) and a remarkably high AUROC (above 0.800 for both variants), reflecting its reliable ability to rank genuinely food-insecure households above food-secure ones.

Moreover, the non-precipitation specification generalises well across time. Its sensitivity experiences an expected, moderate decline of 3.8 p.p., while its accuracy rises by 2.7 p.p. Similarly to the RF, however, the precipitation variant sensitivity declines far more, by 9.5 p.p., representing the steepest CV-to-test sensitivity drop of any specification in this study.

TABLE 6. eXtreme Gradient Boosting: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.395, w/o precip. = 0.400*)

Metric	XGBoost w/ Precip. (CV)	XGBoost w/o Precip. (CV)	XGBoost w/ Precip. (2024 Test)	XGBoost w/o Precip. (2024 Test)
Sensitivity	0.8014	0.8010	0.7060	0.7634
Accuracy	0.6889	0.6843	0.7519	0.7116
Precision	0.2056	0.2030	0.1918	0.1764
False Negative Rate	0.1986	0.1990	0.2940	0.2366
False Positive Rate	0.3228	0.3279	0.2443	0.2927
AUROC	0.8147	0.8127	0.8069	0.8096

Notes: This table presents performance metrics of an eXtreme Gradient Boosting algorithm predicting food-insecure households in PNADC. Columns refer to models with and without precipitation variables assessed via internal cross-validation (2023) and external validation on the unseen 2024 data. Metrics weighted by household survey weight. Abbreviations: XGBoost, eXtreme Gradient Boosting; CV, cross-validation; w/, with; w/o, without; precip., precipitation; AUROC, Area Under the Receiver Operating Characteristic curve.

4.5 SUPPORT VECTOR MACHINES

To further explore the potential of non-parametric methods, we evaluated a polynomial SVM. Rather than partitioning the feature space like the tree-based algorithms discussed above, the SVM identifies the decision boundary that best separates food-insecure from food-secure households, maximising the margin to the nearest training points of each class. The selected second-degree polynomial kernel allows this boundary to curve through the feature space instead of remaining strictly linear, capturing quadratic interactions between predictors rather than relying on a purely linear decision rule (James et al. 2021). Table 7 presents the SVM performance metrics across both the internal 2023 CV and the 2024 external test.

TABLE 7. Support Vector Machine: CV (2023) and final evaluation (2024) performance metrics (*t*: w/precip. = 0.450, w/o precip. = 0.455*)

Metric	SVM w/ Precip. (CV)	SVM w/o Precip. (CV)	SVM w/ Precip. (2024 Test)	SVM w/o Precip. (2024 Test)
Sensitivity	0.7452	0.7312	0.7110	0.7277
Accuracy	0.6765	0.6724	0.6843	0.6717
Precision	0.1903	0.1860	0.1552	0.1522
False Negative Rate	0.2548	0.2688	0.2890	0.2723
False Positive Rate	0.3306	0.3337	0.3179	0.3329
AUROC	0.7762	0.7637	0.7619	0.7565

Notes: This table presents performance metrics of a support vector machine algorithm predicting food-insecure households in PNADC. Columns refer to models with and without precipitation variables assessed via internal cross-validation (2023) and external validation on the unseen 2024 data. Metrics weighted by household survey weight. Abbreviations: SVM, Support Vector Machine; CV, cross-validation; w/, with; w/o, without; precip., precipitation; AUROC, Area Under the Receiver Operating Characteristic curve.

When compared to the other algorithms evaluated, the SVM's metrics remain competitive across the board. Operating at its own optimal thresholds ($t^* = 0.450$ with precipitation and $t^* = 0.455$ without), it reaches a 2024 sensitivity of 71.1 per cent and 72.8 per cent, respectively. Its overall accuracy stands at 68.4 per cent for the precipitation-inclusive variant and 67.2 per

cent for the non-precipitation one. Finally, independent of these thresholds, the model yields AUROC scores of 0.762 and 0.757, demonstrating an overall discriminatory power that is on par with some algorithms in this study.

The SVM variant without rainfall predictors maintains consistent results across temporal contexts. Sensitivity declines by a negligible 0.4 p.p. from CV to the 2024 test, and accuracy stays virtually unchanged. The precipitation-inclusive variant exhibits a different trend: sensitivity drops by 3.4 p.p., a steeper decline, while accuracy marginally increases by 0.8 p.p.

4.6 LASSO LOGISTIC REGRESSION

To bridge the gap between the manually specified LR and the non-parametric ML algorithms, which are trained on the full feature set, we additionally estimate a LASSO logistic regression. This allows us to differentiate whether any performance variations observed in the complex ML algorithms stem from the richer feature information or from their non-linear flexibility.

The results of the LASSO, shown in Table 8, demonstrate a predictive capacity comparable to that of the more complex, non-linear ML approaches. At its optimal threshold ($t^* = 0.440$ for both variants), it reaches a 2024 sensitivity of 75.2 per cent with precipitation and 74.7 per cent without it, ranking among the highest in the study. Overall accuracy stands at 65.7 per cent and 66.3 per cent for the two respective variants, comparable with other algorithms. Finally, the LASSO yields AUROC scores of 0.767 and 0.768 for the precipitation and non-precipitation specifications, respectively.

TABLE 8. LASSO Logistic Regression: CV (2023) and final evaluation (2024) performance metrics (*t^*: w/precip. = 0.440, w/o precip. = 0.440*)

Metric	LASSO w/ Precip. (CV)	LASSO w/o Precip. (CV)	LASSO w/ Precip. (2024 Test)	LASSO w/o Precip. (2024 Test)
Sensitivity	0.7438	0.7448	0.7521	0.7473
Accuracy	0.6639	0.6612	0.6570	0.6627
Precision	0.1838	0.1826	0.1497	0.1513
False Negative Rate	0.2562	0.2552	0.2479	0.2527
False Positive Rate	0.3444	0.3476	0.3509	0.3443
AUROC	0.7692	0.7666	0.7671	0.7682

Notes: This table presents performance metrics of a LASSO logistic regression predicting food-insecure households in PNADC. Columns refer to models with and without precipitation variables assessed via internal cross-validation (2023) and external validation on the unseen 2024 data. Metrics weighted by household survey weight. Abbreviations: CV, cross-validation; w/, with; w/o, without; precip., precipitation; AUROC, Area Under the Receiver Operating Characteristic curve.

Similar to the selected logistic specifications in this study, the LASSO generalises well across time. Sensitivity for the non-precipitation specification rises by a negligible 0.2 p.p. from the internal CV to the 2024 test, while overall accuracy remains nearly unchanged, increasing by just 0.2 p.p. The precipitation-inclusive variant exhibits a similar pattern for sensitivity, rising by 0.8 p.p., though its accuracy moves the other way, declining by 0.7 p.p.

5 AN ASSESSMENT OF THE RESULTS

This section synthesises the evidence across the considered models. We first examine whether including climatic variables improves algorithmic performance. We then compare the operational trade-offs between algorithm families based on their performance metrics. The third subsection analyses the feature selection properties of the LASSO and what they reveal about the importance of algorithmic feature selection. Finally, we characterise how model performance responds to a shared set of standard classification thresholds and what that implies for policy design.

5.1 TO INCLUDE OR NOT CLIMATIC VARIABLES

The results of this study demonstrate that the models trained primarily on socioeconomic and demographic drivers of food insecurity generalise more reliably to the external, unseen test data than those trained with rainfall predictors, with one exception discussed below. Specifically for the RF and SVM, rainfall predictors seemingly provided a marginal predictive signal during the training phase. However, that apparent gain did not survive the 2024 external test. For the LR, RF, SVM, and XGBoost algorithms, incorporating these environmental variables widened the performance gap between the internal CV and the 2024 evaluation significantly more than the baseline non-precipitation specifications.

When comparing the two LR specifications that share all controls but differ solely in the inclusion of pluviometric variables, the non-precipitation specification (M4) yields higher sensitivity than the full rainfall model (M3) at both stages. In CV, this gap is 73.7 per cent vs. 72.9 per cent; in the 2024 test, it is 71.2 per cent vs. 67.9 per cent. The most critical distinction, however, lies in their respective generalisation gaps. M3 suffers a sensitivity decline of 4.9 p.p. between the CV and the 2024 test, which is nearly double the 2.5 p.p. decline observed in M4. This instability shows evidence that precipitation predictors well-fitted to 2023 conditions failed to generalise to the 2024 rainfall patterns for this algorithm. Despite these changes, overall discriminatory power remains virtually unchanged, as M4 reaches an AUROC of 0.766 against M3's 0.759.

RF results extend the same pattern. For the non-precipitation specification, sensitivity declines by 3.4 p.p. (from 77.9 per cent to 74.5 per cent), whereas the precipitation-inclusive variant suffers a steeper 7.0 p.p. (from 79.5 per cent to 72.5 per cent), though AUROC stays almost unchanged between the two variants (0.794 and 0.794). This mirrors the SVM's behaviour: sensitivity drops by a negligible 0.4 p.p. without precipitation (73.1 per cent to 72.8 per cent) compared to a 3.4 p.p. decline with it (74.5 per cent to 71.1 per cent), while the AUROC again remains stable across variants (0.757 and 0.762). XGBoost shows the same dynamic but on a larger scale. Its sensitivity declines by 3.8 p.p. without precipitation (80.1 per cent to 76.3 per cent) and plunges by 9.5 p.p. with it (80.1 per cent to 70.6 per cent), even as its AUROC holds steady (0.810 and 0.807). In every case, the 2024 sensitivity is higher for the non-precipitation variant, although the margin varies: a narrow 1.7 p.p. for the SVM, 2.0 p.p. for the RF, and a substantial 5.7 p.p. for XGBoost.

These patterns observed for the LR, RF, XGBoost, and SVM provide evidence that precipitation data introduces covariate shift[18] without improving policy reach. In all cases, the algorithms learned decision boundaries defined by 2023's specific rainfall distribution,

18. Covariate shift occurs when the distribution of observable characteristics differs between training and application samples, even though the underlying relationship between those characteristics and the outcome of interest remains unchanged (Sugiyama et al. 2007).

which partly broke when 2024 pluviometric conditions differed. Rainfall patterns fluctuate considerably more from year to year than stable, structural household characteristics such as family composition or the household head's education level. This makes a decision boundary anchored to one year's rainfall distribution a fragile basis for prediction in another year, a vulnerability that becomes especially pronounced in the face of extreme climatic shocks.

Despite this instability, AUROC remains almost unchanged across variants for all four algorithms, meaning rainfall barely affects how well these models rank food-insecure households relative to food-secure ones. However, sensitivity is the metric that determines how many food-insecure households a deployed system actually flags at its chosen threshold, and the non-precipitation variants have higher sensitivity in every case, meaning they correctly flag more of the food-insecure households the policy is meant to reach.

The LASSO and the DT break this pattern in different ways, as the addition of rainfall predictors does not destabilise either model. For the LASSO, sensitivity transitions into the 2024 external test either unchanged or slightly higher in both variants, rather than experiencing a decline. The precipitation-inclusive specification rises by 0.8 p.p. (from 74.4 per cent to 75.2 per cent), while the non-precipitation specification remains virtually unchanged, rising by a marginal 0.2 p.p. (from 74.5 per cent to 74.7 per cent). The DT's sensitivity does decline into 2024, but by nearly the same amount whether precipitation is included or not: 4.9 p.p. with it (from 78.4 per cent to 73.5 per cent) and 4.7 p.p. without it (from 78.5 per cent to 73.8 per cent). Furthermore, the AUROC remains essentially static across variants for both algorithms: 0.767 with precipitation and 0.768 without for the LASSO, and 0.760 for both DT variants. This reinforces the finding that rainfall predictors barely affect how well either model ranks food-insecure households against secure ones.

Ultimately, our results suggest that structural and socioeconomic factors, rather than weather, dominate the baseline risk of household food insecurity. The AUROC stays essentially flat across every algorithm's variants regardless of whether precipitation predictors are included, meaning that the overall ranking quality remains unaffected. Extreme weather events remain a real trigger of acute crises, but incorporating a single year's rainfall distribution into a model designed for policy targeting must be approached with caution. Four of the six algorithms evaluated lost operational sensitivity when tested against a rainfall year on which they had not trained. This underscores why our two-stage evaluation methodology, with its strict temporal separation, serves as an essential safeguard to prevent the deployment of inconsistent algorithms into policymaking. A definitive conclusion on this issue, however, requires data spanning multiple training and testing years.

Anecdotal evidence points to a possible explanation for the declining performance from the 2023 internal validation to the 2024 external one. Heavy rainfall followed by widespread severe floods in *Rio Grande do Sul*, one of the most economically advantaged states in Brazil, beginning in late April 2024, may be behind this drop. In 2023, the highest levels of rainfall were concentrated in remote areas of the Amazon region, one of the poorest in Brazil. Then the shift to the rich South, which also counted on significant policy response to the floods, could be a confounder in the model. This could be a hypothesis to be tested in future work.

Contrary to the findings presented here, previous research by our team had indicated that households experiencing the lowest or the highest levels of precipitation in the months preceding the survey would have the highest risks of falling into food insecurity (Campante Vale 2026). The results of the present study do not rule out the earlier findings in their static interpretation. That is: everything else constant, residents of more arid or more humid zones

of Brazil are those experiencing higher levels of food insecurity on average. Nevertheless, a plausible explanation is that the very arid or very humid places that may be driving these results, such as the Northeast hinterlands or the inner Amazon tropical forest, are chronically poor and, thus, marked by low indicators of socioeconomic developments already captured on other variables. Adding climatic variables that capture variations in other regions may only add noise to the estimates. In forthcoming work, we plan to delve further into this discussion.

It is important to note that our current work is limited by its specific setting. We have worked using aggregation of precipitation metrics in large areas – the best we can do with PNADC data—and standardising the relevant time span over this continental-sized country. Models that can apply more granular data, create shock indicators from historical patterns and adjust relevant rainfall periods to local crops seasons may perform better. To definitively conclude that climatic data is not useful to predict food insecurity in Brazil, other alternative features commonly tested in the literature, such as temperature anomalies or drought indices, must be tested.

5.2 THE TRADE-OFFS OF MODEL SELECTION

No algorithm in this comparison wins on every metric. At its own optimal threshold, each model trades sensitivity, the share of food-insecure households it actually flags, against accuracy, precision, and the FP rate it generates in exchange. Table 9 puts these trade-offs side by side for the non-precipitation variant of each algorithm and the M2 specification of the LR (which includes demographic, labour market, and rainfall predictors).

TABLE 9. Model performance metrics (2024 Test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2)

Metric	LR (M2)	DT	RF	XGBoost	SVM	LASSO LR
Sensitivity	0.763	0.737	0.745	0.763	0.728	0.747
	(0.0058)	(0.0061)	(0.0059)	(0.0057)	(0.0059)	(0.0057)
Accuracy	0.637	0.675	0.722	0.712	0.672	0.663
	(0.0017)	(0.0017)	(0.0016)	(0.0017)	(0.0017)	(0.0017)
Precision	0.144	0.155	0.179	0.176	0.152	0.151
	(0.0020)	(0.0020)	(0.0024)	(0.0024)	(0.0021)	(0.0021)
False Negative Rate	0.237	0.263	0.255	0.237	0.272	0.253
	(0.0058)	(0.0061)	(0.0059)	(0.0057)	(0.0059)	(0.0057)
False Positive Rate	0.373	0.330	0.280	0.293	0.333	0.344
	(0.0018)	(0.0018)	(0.0017)	(0.0017)	(0.0018)	(0.0018)
AUROC	0.7581	0.7596	0.7938	0.8096	0.7565	0.7682
t^*	0.475	0.385	0.225	0.400	0.455	0.440

Notes: This table presents the performance metrics across non-precipitation ML model variants and the M2 LR, assessed via external validation on unseen data (2024). Metrics weighted by household survey weight. Bootstrap standard errors for each metric are shown in parentheses (B = 2000 household-level resamples). Abbreviations: LR, logistic regression; DT, decision tree; RF, random forest; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; LASSO LR, Least Absolute Shrinkage and Selection Operator Logistic Regression; AUROC, Area Under the Receiver Operating Characteristic curve.

For the policy objective of effectively targeting the highest share of truly food-insecure households, sensitivity is the most consequential metric. The selected LR specification (M2) and XGBoost algorithm tie as the top performers in this regard, each at 76.3 per cent, indicating that either model would correctly identify roughly three of every four households genuinely at risk. However, the respective trade-offs required by each algorithm to achieve this high sensitivity differ considerably.

The LR achieves this sensitivity at the expense of nearly every other metric. Among the six models evaluated, it records the highest FP rate (37.3 per cent), alongside the lowest accuracy (63.7 per cent) and precision (14.4 per cent). Additionally, its performance is highly sensitive to the chosen specification. On the 2024 test alone, sensitivity across the four logistic specifications ranges from 67.9 per cent (M3) to 76.3 per cent (M2), while accuracy moves in the opposite direction, from 70.5 per cent (M3) to 63.7 per cent (M2). Consequently, the specification yielding the highest sensitivity simultaneously produces the lowest accuracy, and vice versa. This trade-off confirms that the LR's efficacy depends on manual predictor selection. The LASSO offers a route around this dependence through automatic feature selection, which is further discussed in the following subsection.

This trade-off reflects a fundamental statistical mechanism, rather than a limitation specific to the LR. Any specification or model that classifies a larger proportion of households as at-risk effectively casts a wider net, successfully capturing more truly food-insecure households (increasing sensitivity), but also flags more food-secure households alongside them. As food-insecure households represent a minority of the 2024 sample, these new FP lower precision and consequently depress overall accuracy.

The various LR specifications cast nets of different widths depending on which predictors they include, which explains why sensitivity and accuracy move in opposite directions across them. This pattern operates within every algorithm evaluated in this study; what differs is how cleanly each one separates the two classes, and therefore how much accuracy a wider net actually costs. Essentially, our aim throughout this analysis is to find an algorithm that reaches more of the at-risk population without a proportional rise in that cost, striking a more balanced trade-off than the LR.

XGBoost offers this more balanced alternative, emerging as the best-performing model of our exercise. Besides achieving the highest sensitivity, its accuracy and precision are also among the highest in the comparison, achieving a balance that is matched only by the RF. Its non-precipitation variant matches the LR M2, correctly flagging three of every four truly food-insecure households, but requires fewer performance trade-offs. It sustains an overall accuracy of 71.2 per cent against the logit's 63.7 per cent (correctly classifying about seven in ten households overall), alongside a precision of 17.6 per cent against the logit's 14.4 per cent. In practical terms, this means approximately one in six households it flags as at-risk are genuinely food-insecure, compared to one in seven for the LR.

XGBoost's AUROC (0.810) also stands apart from the rest of the field. Only the RF performs comparably, while the remaining four models cluster well behind. Expressed intuitively, if given one food-insecure and one food-secure household at random, the model correctly assigns a higher risk probability to the food-insecure household 81 per cent of the time. Finally, one last advantage that should be highlighted is the fact that XGBoost trained in a fraction of the time required by the other ensemble method in our application, an expected outcome given that the algorithm is engineered for computational efficiency (Chen and Guestrin 2016).

The RF emerges as XGBoost's closest competitor, achieving its own robust balance that skews

slightly further toward precision than broad reach. It has the highest accuracy (72.2 per cent) and precision (17.9 per cent) of the six models, alongside the lowest FP rate (28.0 per cent). To obtain this performance, it exchanges a fraction of XGBoost's sensitivity, though it still correctly flags a strong 74.5 per cent of all truly food-insecure households. Its AUROC (0.794) ranks second overall and sits nearly on par with XGBoost's, indicating it separates food-insecure from food-secure households almost as reliably as the top-performing model. However, its primary drawback is computational, as the process of growing and tuning a large ensemble of trees made it by far the most time-intensive model to train in our application, taking roughly five times as long as XGBoost.

The remaining algorithms are also competitive, with similar performances that fall just a tier below XGBoost and the RF. First, the DT reaches 73.8 per cent sensitivity and 67.5 per cent accuracy, yielding an AUROC (0.760) virtually identical to that of the LR baselines. As the simplest ML algorithm in this comparison, it offers a distinct practical advantage: whereas the RF, XGBoost, and SVM all require some form of hyperparameter search, the DT is pruned directly from a single fitted tree, which makes it the cheapest and fastest of the four ML algorithms to train. This is a viable option for deployments facing restricted computing infrastructure.

Second, the SVM trails the other models by a fractional margin, recording the lowest sensitivity (72.8 per cent) and the lowest AUROC (0.757) of the six, though these results reflect the capped training regime it operates under, which may understate its true potential (see Appendix A for more information). Despite this, its remaining metrics are highly competitive: its accuracy (67.2 per cent) exceeds both the LR M2 (63.7 per cent) and the LASSO (66.3 per cent), while its precision (15.2 per cent) matches the LASSO and surpasses the LR (14.4 per cent). Finally, the LASSO performs at a comparable, though marginally superior, level (74.7 per cent sensitivity, AUROC of 0.768). Because its automatic feature selection directly addresses the specification dependencies observed in the LR, its results are examined in detail in the following subsection.

5.3 FEATURE SELECTION OR ALGORITHMIC FLEXIBILITY: INSIGHTS FROM THE LASSO

A central methodological question motivating the inclusion of the LASSO in our analysis is to assess whether the performance variations observed in the non-parametric approaches, compared to the LR baselines, are driven by access to a richer feature set or by the non-linear flexibility of the algorithms themselves. The LASSO is fundamentally an extension of the LR, sharing its linear, parametric functional form but replacing its manually chosen specification with an L1 penalty that allows the model to process the full feature space and automatically select which predictors matter.

Under this regularisation, the coefficients of less informative variables are shrunk toward zero, while the least relevant are driven to exactly zero, effectively removing them from the model (Hastie et al. 2009). Because the LASSO and the standard LR share this structural foundation, any performance gap between them can be attributed strictly to the effect of automatic feature selection. This makes the LASSO an ideal bridge to isolate this specific effect from the potential performance gains that arise from non-linearity.

When compared to the selected LR specification (M2), the LASSO trades a marginal drop in sensitivity of just 1.6 percentage points (74.7 per cent against 76.3 per cent) for gains across every other metric. It achieves a higher accuracy (66.3 per cent against 63.7 per cent), higher precision (15.1 per cent against 14.4 per cent), a lower FP rate (34.4 per cent against 37.3 per cent), and a higher AUROC (0.768 against 0.758). Its main advantage, though, is structural

rather than purely predictive: automatic feature selection removes the need to manually choose a specification in advance. This choice may be inherently unstable, as evidenced by the LR's 2024 sensitivity swinging by 8.4 percentage points (from 67.9 per cent to 76.3 per cent) depending solely on which predictors are included. Because it resolves this specification vulnerability while performing on par with the optimal LR baseline, the LASSO stands as a highly viable methodological alternative for this targeting problem in its own right.

When compared to the non-linear ML models, the LASSO remains competitive across every metric evaluated. Its 2024 sensitivity (74.7 per cent) marginally exceeds that of the DT (73.8 per cent), the RF (74.5 per cent), and the SVM (72.8 per cent), trailing only XGBoost (76.3 per cent). Also, its accuracy (66.3 per cent) and precision (15.1 per cent) closely track the performance of the DT (67.5 per cent and 15.5 per cent) and the SVM (67.2 per cent and 15.2 per cent). In terms of overall discriminatory power, its AUROC (0.768) outperforms both the DT (0.760) and the SVM (0.757), though it trails the RF (0.794) and XGBoost (0.810). Taken together, the LASSO proves competitive with the DT, edges ahead of the SVM, and falls behind only the two ensembles.

These findings have important implications for applied policy research. First, a substantial portion of the achievable predictive performance is already captured by a regularised linear model. However, the small gains derived from algorithmic complexity are not trivial: the RF and XGBoost improve accuracy over the LASSO by 5.9 and 4.9 percentage points respectively (72.2 per cent and 71.2 per cent against 66.3 per cent), and they rank food-insecure households above food-secure ones more reliably, with higher AUROCs. Consequently, we can conclude that, in this context, increased algorithmic complexity does yield concrete performance gains.

Nevertheless, these results must be interpreted with caution. In public policy contexts where institutional transparency and model interpretability are paramount, the LASSO remains a highly credible alternative. Because it provides interpretable coefficients while bypassing the vulnerabilities of manual specification through automatic feature selection, it remains a reliable and stable option to the traditional LR (Hastie et al. 2009). In addition, the researcher must bear in mind that highly flexible models are prone to overfitting to training sample idiosyncrasies, necessitating a carefully designed evaluation strategy. In our case, cross-validation and the external testing mitigate this risk, though attention remains necessary when making further extrapolations.

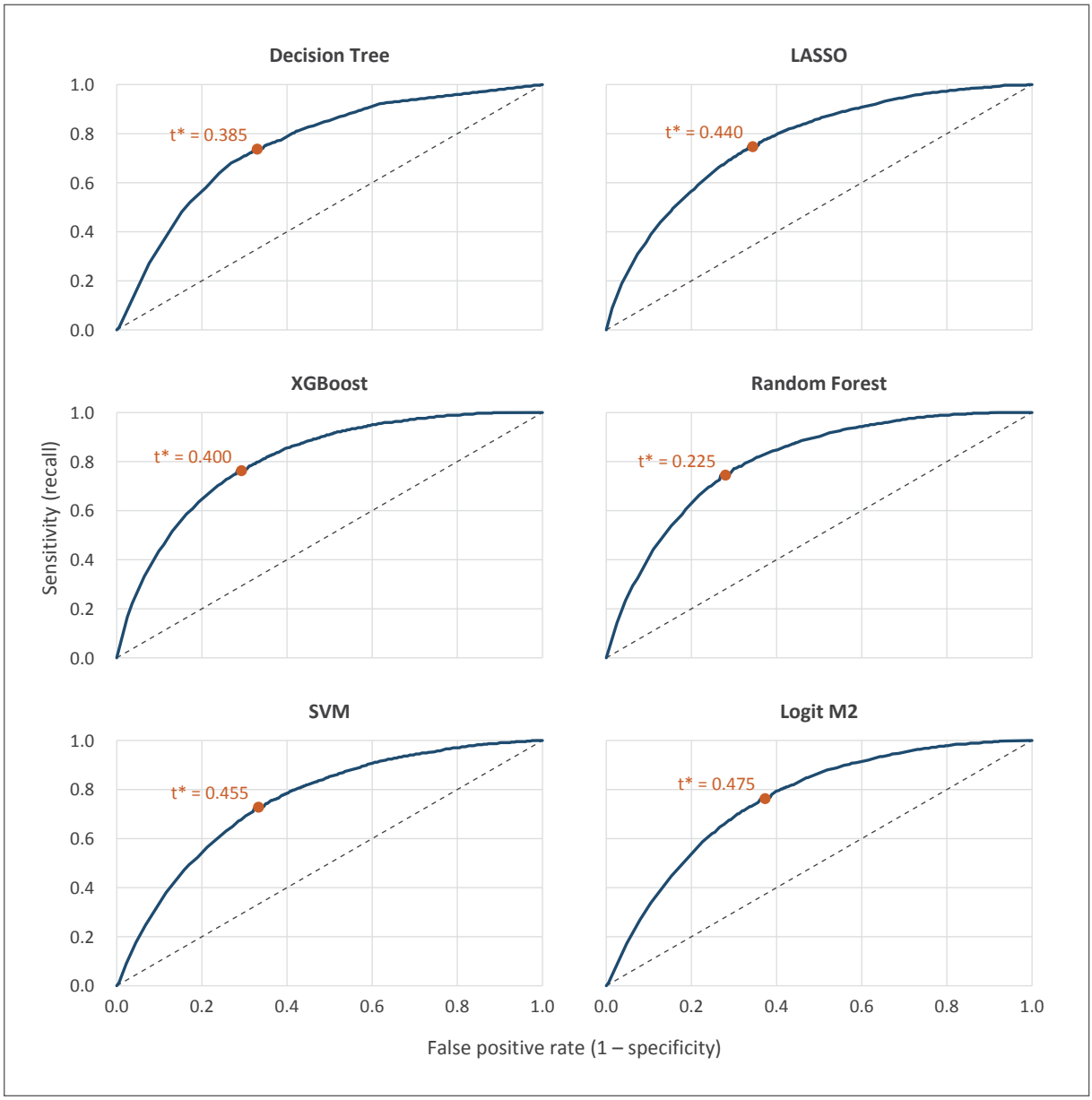
In our setting, we compared feature importance across both LASSO and more complex models. It was reassuring to find that the variables with the highest weights assigned by the algorithms aligned with theoretical expectations, with household income ranking at the top. While this does not replace out-of-sample validation, checking for face validity across models is a crucial step in ensuring the structural logic of the forecast.

5.4 THRESHOLD CHOICE AND POLICY TRADE-OFFS

Maximising sensitivity has been the primary criterion throughout this analysis, since broad reach is the fundamental objective of an early-warning system designed to detect households at risk of food insecurity. Precision, on the other hand, captures a complementary dimension: the share of flagged households that are genuinely food insecure. Both metrics are inextricably connected through the FP rate. Casting a wider net to catch more truly food-insecure households raises sensitivity, but it inevitably flags more food-secure households alongside them, and every one of those FP subsequently dilutes precision.

Crucially, this trade-off is not static, as it depends on where the classification threshold is set. Each model's optimal operating threshold (t^*) was chosen by maximising Youden's J, the criterion that finds the point on the model's ROC curve that optimally balances sensitivity against the FP rate. This approach yields a mathematically defined threshold rather than an arbitrary one, and, consequently, the same rule produces a different t^* for each model. Figure 1 illustrates this calibration in practice, plotting each model's ROC curve for the 2024 external test in its own panel, with its respective CV-optimal t^* marked as the operating point.

FIGURE 1. ROC curves (2024 test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2), with each model's CV-optimal threshold (t^*) marked

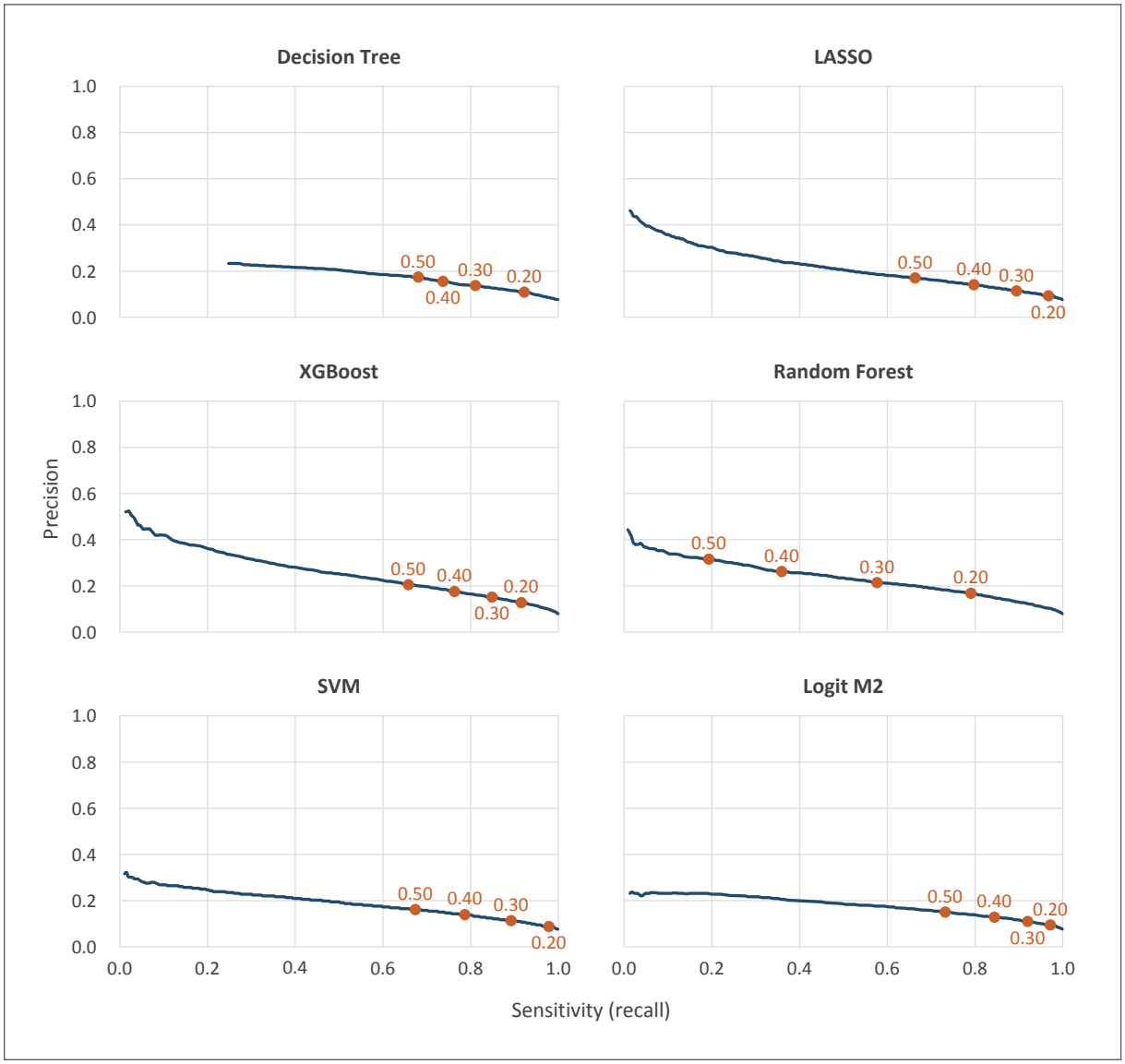


Notes: This figure presents the ROC curves of the non-precipitation ML model variants and the *M2* LR, assessed via external validation on unseen data (2024). The x-axis represents the FP rate (1 minus specificity), and the y-axis represents sensitivity (recall). Highlighted markers indicate each model's own classification threshold (t^*), calibrated via CV and applied unchanged to this external test. Metrics weighted by household survey weight. Abbreviations: LR, logistic regression; DT, decision tree; RF, random forest; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; LASSO LR, Least Absolute Shrinkage and Selection Operator Logistic Regression; CV, cross-validation.

Source: Author's elaboration.

We mark the 2023-derived t^* on the 2024 ROC curve to see where it lands on the unseen data distribution, mirroring a real deployment of the tool to policymaking. If it still sits in a well-balanced region, that is evidence the calibration generalises to new data. Read side by side, the panels make the case for letting each algorithm set its own threshold. Because each model's ROC curve has a slightly different shape, applying the same calibration rule to six different distributions naturally produces six different t^* values, from 0.225 to 0.475. Fixing a common threshold would land at a different point on every curve that may not be the most balanced one for each specific model. RF illustrates this clearly: its operating point sits at $t^* = 0.225$, well below the 0.385–0.475 band occupied by the other five models.

FIGURE 2. Precision-recall curves (2024 test). Non-precipitation ML model variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2)



Notes: This figure presents the precision-recall curves of the non-precipitation ML model variants and the *M2* LR, assessed via external validation on unseen data (2024). The x-axis represents sensitivity (recall), and the y-axis represents precision. Highlighted markers indicate specific operating points evaluated at classification thresholds of 0.20, 0.30, 0.40, and 0.50. Metrics weighted by household survey weight. Abbreviations: LR, logistic regression; DT, decision tree; RF, random forest; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; LASSO LR, Least Absolute Shrinkage and Selection Operator Logistic Regression.

Source: Author's elaboration.

In practice, however, a policymaker deploying one of these models rarely has the luxury of letting the algorithm unilaterally dictate its own threshold, as non-statistical factors must be accounted for. Threshold choice is, fundamentally, a policy decision. For a targeted social protection programme, it mirrors the trade-off policymakers face between broadening coverage of the eligible population and containing the fiscal cost of misclassification. A lower threshold identifies more food-insecure households but demands a larger budget, as it simultaneously expands coverage and tolerates more false flags. This leniency may be entirely deliberate, as the human cost of missing a genuinely vulnerable household often outweighs the fiscal cost of serving a few who are not.

To observe how sensitivity, precision, and accuracy respond to these policy choices, we evaluated each model's performance across a range of standard thresholds (0.20 to 0.50). This also allowed us to test whether the favourable balance achieved by XGBoost and RF at their own t^* extends to thresholds neither model was calibrated for. Figure 2 maps the precision–recall (PR) curves for all six models on the 2024 external test, plotting precision against recall (sensitivity) as the threshold sweeps across its full range, using the non-precipitation ML model variants alongside the M2 specification for the LR. The four labelled points on each curve mark the selected thresholds.

Because food-insecure households represent only about 9.4 per cent of the survey-weighted sample, even a modest FP rate among the much larger food-secure majority produces enough FP to rival the true positives. This demographic configuration of the data places a ceiling on precision for every model regardless of its discriminative power, but this ceiling still differs sharply across models. XGBoost's is the highest (approximately 0.52), followed by LASSO (approximately 0.46) and RF (approximately 0.44), though all three reach these at very low sensitivity (recall). SVM's ceiling is noticeably lower (approximately 0.33), and the DT and LR M2 trail furthest behind, never exceeding roughly 0.23 at any point on their curves.

The PR curves for XGBoost, RF, and LASSO all run continuously across nearly the full sensitivity range, from near-zero out to exhaustive coverage. The logit and SVM curves span this same range but sit at consistently lower precision throughout. Consequently, matching the reach of XGBoost, RF, or LASSO at any given point means accepting a steeper FP cost for both algorithms. The DT curve stands apart from the other five, covering a narrower band of the PR space. Table 10 complements this analysis by detailing performance metrics across the selected threshold range for the non-precipitation DT, LASSO, XGBoost, RF, SVM, and the selected logistic M2 specification.

The performance trade-off is steep for every model. As the threshold lowers from 0.50 to 0.20, the DT shifts modestly: sensitivity rises from 68.0 per cent to 92.2 per cent (+24.2 p.p.), while accuracy falls from 72.9 per cent to 42.3 per cent (–30.6 p.p.) and the FP rate climbs from 26.7 per cent to 61.9 per cent (+35.2 p.p.). The logit shows a similar sensitivity gain, from 73.3 per cent to 97.2 per cent (+23.9 p.p.), but experiences more severe swings in both accuracy (–37.8 p.p.) and the FP rate (+42.8 p.p.). The LASSO swings even further, trading a considerable 44.5 p.p. drop in accuracy and a 50.6 p.p. spike in FP to increase its sensitivity by more, from 66.4 per cent to 96.8 per cent (+30.4 p.p.). The SVM represents the most extreme case, securing a comparable sensitivity gain relative to LASSO (+30.5 p.p.) at the cost of the largest variance in both accuracy (–48.6 p.p.) and FP rate (+55.1 p.p.).

TABLE 10. Threshold sensitivity analysis (2024 Test). Non-precipitation ML variants (DT, LASSO, XGBoost, RF, SVM) and LR (M2)

Model	Threshold	Sensitivity	Accuracy	Precision	FPR	FNR
DT	0.2	0.922	0.423	0.109	0.619	0.078
DT	0.3	0.811	0.595	0.136	0.422	0.189
DT	0.4	0.737	0.675	0.155	0.330	0.263
DT	0.5	0.680	0.729	0.173	0.267	0.320
LASSO	0.2	0.968	0.285	0.093	0.771	0.032
LASSO	0.3	0.895	0.464	0.114	0.571	0.105
LASSO	0.4	0.798	0.614	0.140	0.401	0.202
LASSO	0.5	0.664	0.730	0.171	0.265	0.336
XGBoost	0.2	0.916	0.521	0.128	0.511	0.084
XGBoost	0.3	0.849	0.626	0.151	0.392	0.151
XGBoost	0.4	0.763	0.712	0.176	0.293	0.237
XGBoost	0.5	0.658	0.782	0.206	0.208	0.342
RF	0.2	0.791	0.684	0.167	0.325	0.209
RF	0.3	0.577	0.807	0.214	0.174	0.423
RF	0.4	0.359	0.875	0.262	0.083	0.641
RF	0.5	0.193	0.907	0.314	0.035	0.807
SVM	0.2	0.979	0.224	0.088	0.838	0.021
SVM	0.3	0.893	0.461	0.113	0.574	0.107
SVM	0.4	0.786	0.613	0.139	0.401	0.214
SVM	0.5	0.674	0.710	0.162	0.287	0.326
LR (*M2*)	0.2	0.972	0.287	0.094	0.769	0.028
LR (*M2*)	0.3	0.921	0.424	0.109	0.616	0.079
LR (*M2*)	0.4	0.845	0.556	0.129	0.467	0.155
LR (*M2*)	0.5	0.733	0.665	0.150	0.341	0.267

Notes: This table presents the performance metrics of the non-precipitation ML model variants and the *M2* LR, assessed via external validation on unseen data (2024) across different classification thresholds: 0.20, 0.30, 0.40, and 0.50. Metrics weighted by household survey weight. Abbreviations: LR, logistic regression; DT, decision tree; RF, random forest; XGBoost, eXtreme Gradient Boosting; SVM, support vector machine; LASSO LR, Least Absolute Shrinkage and Selection Operator Logistic Regression; FPR, False Positive Rate; FNR, False Negative Rate.

RF stands as an outlier. As Figure 2 shows, its four threshold markers are spread out across the PR curve rather than clustered together like the other algorithms. As the threshold declines from $t = 0.50$ to $t = 0.20$, RF's sensitivity surges from 19.3 per cent to 79.1 per cent (+59.8 p.p.), while its accuracy falls from 90.7 per cent to 68.4 per cent (-22.3 p.p.) and the FP rate rises from 3.5 per cent to 32.5 per cent (+29.0 p.p.). For comparison, the DT and SVM gain only 24.2 p.p. and 30.5 p.p. of sensitivity over this exact same range, respectively, yet they pay a much higher penalty in both accuracy and FP rate.

This asymmetry reveals a crucial operational pattern: RF "buys" each point of sensitivity more cheaply than the other five models in terms of the accuracy trade-off. Its cost per point of sensitivity gained, roughly a third of a percentage point of accuracy and half a percentage point of FP rate, is three to four times cheaper than the other algorithms' average in this dimension, and still roughly two and a half times cheaper than XGBoost, the closest of the five on this measure. Within the standard 0.20–0.50 threshold range, RF has the highest accuracy

and precision among all the tested models. However, within this same range, RF never reaches the levels of targeting population's coverage the other five do. At $t = 0.20$, its 79.1 per cent sensitivity falls well short of the near-exhaustive reach the other models achieve at that same threshold (91.6 per cent–97.9 per cent), only matching them when they operate at their own more conservative $t = 0.40$ (73.7–84.5 per cent).

XGBoost's own calibrated threshold ($t^* = 0.40$) falls exactly on this standard grid. Across the full 0.20–0.50 range, XGBoost's sensitivity moves from 65.8 per cent to 91.6 per cent (+25.8 p.p.), its accuracy from 78.2 per cent to 52.1 per cent (-26.1 p.p.), and its FP rate from 20.8 per cent to 51.1 per cent (+30.3 p.p.). For this algorithm, each point of sensitivity costs roughly one point of accuracy and 1.2 points of FP rate, the smallest cost of any model besides RF, confirming that XGBoost's strong balance is not merely an artefact of its specific t^*.

Fundamentally, the PR curves and threshold sensitivity analysis demonstrate that no single model strictly dominates across every dimension. RF has the most efficient sensitivity-per-false-positive-rate ratio of the six algorithms, yet it fails to reach the high-sensitivity range the other five occupy, which is required for broad preventative coverage. Among the models that do achieve the necessary reach, XGBoost offers the superior balance, consistent with its threshold-free AUROC.

Which model is preferred depends on what the policy exercise prioritises. For an early-warning tool built to reach as many at-risk households as possible, sensitivity is what matters most, and on that basis XGBoost is the best choice. It marginally outperforms the RF at its calibrated threshold, and crucially, it keeps climbing toward 91.6 per cent sensitivity while the RF hits a functional ceiling of 79.1 per cent within our threshold grid. A programme equipped with the fiscal capacity to serve a broad population can effectively pair XGBoost with a low threshold. Conversely, a programme forced to ration its coverage more may prefer RF's efficiency stemming from its high accuracy and precision, or simply choose a higher threshold to any of the six algorithms reducing the false positive rates. When there is flexibility enough to balance between sensitivity and specificity, models with Youden's J calibrated thresholds delivered qualitatively similar outcomes. However, the right combination changes as the programme's targeting priorities and budget shift, so threshold calibration must be treated as a dynamic policy lever rather than a fixed statistical benchmark.

6 POLICY DISCUSSION

Before implementing any predictive model, policymakers must clearly define their ultimate objective. This choice is not merely technical, it fundamentally determines how models should be designed, evaluated, and deployed. This work was largely inspired by the pioneering initiative of CadINSAN, which uses *PNAD Contínua* data to estimate the probability of food insecurity and extrapolates it to Brazil's social registry (*Cadastro Único*). According to its methodological report, CadINSAN aims to monitor the magnitude and incidence of food insecurity risk at the municipal level to inform policy decisions (MDS 2025).

While this study shares the goal of supporting policymaking, we argue that food insecurity monitoring and early warning are distinct tasks, with different operational implications. Monitoring focuses on describing the current situation, which is usually done by estimating prevalence rates or counting the number of food-insecure households across regions. Early

warning, in contrast, aims to identify which specific households are at risk, enabling timely and targeted intervention. Although more complex, we argue that advancing toward an early-warning system is both feasible and desirable given Brazil's current institutional and data infrastructure.

Food insecurity monitoring operates primarily at the aggregate level. Classification models can support this task by predicting food insecurity in periods or areas where direct measurement is unavailable. For example, Brazil collected EBIA data in the PNADC in 2023 and 2024, and also in the 2024–2025 Household Budget Survey (POF), but, at the time of publication, no such collection had been announced for 2026. In this context, models trained on past surveys could be used to predict food insecurity in new PNADC waves, allowing policymakers to estimate national or regional prevalence rates even in the absence of direct data collection. CadINSAN addresses the spatial gaps by extrapolating PNADC-based models to *Cadastro Único*, producing municipal-level estimates where survey representativeness is limited.

However, monitoring alone is insufficient for effective prevention. Early-warning systems operate at the micro level, seeking to flag households that are likely experiencing, or are about to experience, food insecurity. While aggregate statistics are essential for keeping food insecurity on the policy agenda and guiding resource allocation, only household-level identification allows for timely and targeted responses. This distinction is particularly relevant in large urban areas, where prevalence rates may be relatively low but the absolute number of vulnerable households remains substantial. Importantly, early-warning systems can also generate aggregate indicators; thus, they encompass and extend monitoring rather than replace it.

To successfully operationalise such an optimised, dynamic tool, a highly integrated institutional data infrastructure is essential. Fortunately, Brazil is well positioned to advance toward a national early-warning system. The *Cadastro Único* already provides a comprehensive registry of low-income households, covering 45 per cent of the national population.[19] The reintroduction of EBIA into nationally representative surveys offers a reliable ground truth for model training and validation. In addition, the TRIA (Food Insecurity Risk Screening) food insecurity indicator recorded from primary health care services is increasingly being integrated into the social registry. While the TRIA data may be subject to selection bias and may not be suitable as standalone inputs for prediction models, it offers valuable opportunities for external validation and stress-testing of models trained on survey data.

Importantly, the near-parity in performance between the LASSO and several non-parametric ML algorithms demonstrates that effective early-warning systems do not necessarily need to rely on opaque, complex models. This is particularly relevant for institutional adoption, as more interpretable models facilitate independent auditing, methodological validation, and transparent policy communication. Nevertheless, where technical capacity allows, the deployment of complex ML algorithms remains justified in its own way. XGBoost emerges as the best performing model by sustaining a robust balance across metrics, without favouring reach over targeting efficiency, or vice versa.

Importantly, the development of predictive systems should not be seen as a substitute for

19. Estimated from administrative data available on https://aplicacoes.cidadania.gov.br/vis/data3/data-explorer.php and https://www.gov.br/secom/pt-br/acompanhe-a-secom/noticias/2025/08/populacao-do-brasil-alcanca-marca-de-213-4-milhoes-de-habitantes-divulga-ibge. Accessed on 17 March 2026.

fieldwork or local groups working on the cause. On the contrary, the knowledge and judgment of frontline social workers are invaluable, while civil society vigilance and engagement with policy-decisions remain essential to guarantee access to adequate and nutritious food for those most in need (CAISAN 2025). Much less should the predictions be understood as the determining criterion for who will receive support. Predictive models should be understood as complementary tools, helping to prioritise attention and allocate scarce resources more effectively. This is particularly relevant in contexts where social assistance systems are overstretched and where the most vulnerable households may be precisely those least likely to actively seek support.

As practical suggestions, models could be trained on PNADC itself and the predictions extrapolated to *Cadastro Único* administrative data using the set of compatible variables. This follows the same spirit of the CadINSAN indicator but more ambitiously. A pilot could be run with field validation in a few selected municipalities with small populations, where social workers could visit households to check the predictive performance in-person. Conversely, another initiative would be crossing PNADC predictions with TRIA results, in municipalities where the latter indicator already has a good coverage of CadÚnico families. Later on, with TRIA expansion, a natural extension would be entirely training and testing the models with full CadÚnico data and validating them in the field with the social workers.

From an operational perspective, implementing a first-generation classification system is relatively low-cost, given existing data and institutional capacity. A more advanced early-warning system would require integrating micro-level predictions with macro-level monitoring and incorporating information on shocks (such as economic downturns or localised climatic events) that may trigger transitions into food insecurity. Moving in this direction would allow Brazil not only to measure food insecurity, but to anticipate it and act proactively, strengthening the country's capacity to protect vulnerable populations.

7 REFERENCES

Bakhtsiyarava, M., T. G. Williams, A. Verdin, and S. D. Guikema. 2021. "A Nonparametric Analysis of Household-Level Food Insecurity and Its Determinant Factors: Exploratory Study in Ethiopia and Nigeria." Food Security 13: 55–70. https://doi.org/10.1007/s12571-020-01132-w.

Barbosa, R. M., and D. R. Nelson. 2016. "The Use of Support Vector Machine to Analyze Food Security in a Region of Brazil." *Applied Artificial Intelligence* 30 (4): 318–330. https://doi.org/10.1080/08839514.2016.1169048.

Câmara Interministerial de Segurança Alimentar e Nutricional (CAISAN). 2025. *Balanço de atividades da Câmara Interministerial de Segurança Alimentar e Nutricional (Caisan): a retomada do Sisan e os avanços na garantia do direito humano à alimentação adequada, biênio 2023–2024*. Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome. Brasília, DF.

Campante Vale, R. C. 2026. "Faça chuva ou faça sol: precipitação e insegurança alimentar nos domicílios brasileiros." *Cadernos de Estudos Desenvolvimento Social em Debate* 41: 78–88.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Cherol, C. C. S., A. A. Ferreira, J. B. Lignani, and R. Salles-Costa. 2022. "Regional and Social Inequalities in Food Insecurity in Brazil, 2013–2018." *Cadernos de Saúde Pública* 38: e00083822.

FAO, IFAD, UNICEF, WFP, and WHO. 2022. *The State of Food Security and Nutrition in the World 2022: Repurposing Food and Agricultural Policies to Make Healthy Diets More Affordable*. Rome: FAO. https://doi.org/10.4060/cc0639en.

FAO, IFAD, UNICEF, WFP, and WHO. 2025. *The State of Food Security and Nutrition in the World 2025: Addressing High Food Price Inflation for Food Security and Nutrition*. Rome: FAO. https://doi.org/10.4060/cd6008en.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hoffmann, R. 2008. "Determinantes da insegurança alimentar no Brasil: análise dos dados da PNAD de 2004." *Segurança Alimentar e Nutricional* 15 (1): 49–61.

Hoffmann, R. 2013. "Determinantes da insegurança alimentar no Brasil em 2004 e 2009." *Segurança Alimentar e Nutricional* 20 (2): 219–235.

Hoffmann, R. 2021. "Insegurança alimentar no Brasil após crise, sua evolução de 2004 a 2017–2018 e comparação com a variação da pobreza." *Segurança Alimentar e Nutricional* 28: 1–17.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. Springer. https://doi.org/10.1007/978-1-0716-1418-1.

Janssen, S. M. W., Y. Bouzembrak, and B. Tekinerdogan. 2024. "Artificial Intelligence in Malnutrition: A Systematic Literature Review." *Advances in Nutrition* 15 (9): 100264. https://doi.org/10.1016/j.advnut.2024.100264.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–495. https://doi.org/10.1257/aer.p20151023.

MDS. 2025. *CadInsan: Indicador de Risco de Insegurança Alimentar Municipal a Partir dos Dados do CadÚnico*. Ministério do Desenvolvimento e Assistência Social, Família e Combate à Fome. Secretaria Extraordinária de Combate à Pobreza e à Fome (SECF). Brasília-DF.

Mullainathan, S., and J. Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106. https://doi.org/10.1257/jep.31.2.87.

Platt, J. C. 1999. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods." In *Advances in Large Margin Classifiers*, edited by A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, 61–74. MIT Press.

Pontes, R. P., M. N. Barbosa, C. A. Oliveira, and P. R. Abdallah. 2018. "Quem passa fome no Brasil? Uma análise regional dos determinantes da insegurança alimentar forte nos domicílios brasileiros." *Revista Brasileira de Estudos Regionais e Urbanos* 12 (2): 225–241.

Rao, B., M. Rashid, M. G. Hasan, and G. Thunga. 2025. "Machine Learning in Predicting Child Malnutrition: A Meta-Analysis of Demographic and Health Surveys Data." *International Journal of Environmental Research and Public Health* 22 (3): 449. https://doi.org/10.3390/ijerph22030449.

Rede PENSSAN. 2022. *II Inquérito Nacional sobre Insegurança Alimentar no Contexto da Pandemia da Covid-19 no Brasil*. São Paulo: Fundação Friedrich Ebert.

Reichenheim, M. E., G. S. Interlenghi, C. L. Moraes, A. M. Segall-Corrêa, R. Pérez-Escamilla, and R. Salles-Costa. 2016. "A Model-Based Approach to Identify Classes and Respective Cutoffs of the Brazilian Household Food Insecurity Measurement Scale." *The Journal of Nutrition* 146 (7): 1356–1364.

Ribeiro-Silva, R. de C., M. Pereira, T. Campello, A. J. F. Ferreira, S. M. C. dos Santos, E. Aragão, J. M. de M. Guimarães, and M. L. Barreto. 2020. "Implicações da pandemia COVID-19 para a segurança alimentar e nutricional no Brasil." *Ciência & Saúde Coletiva* 25 (9): 3421–3430. https://doi.org/10.1590/1413-81232020259.22152020.

Saihood, Q., and E. Sonuç. 2022. "The Efficiency of Classification Techniques in Predicting Anemia among Children: A Comparative Study." In *Emerging Technology Trends in Internet of Things and Computing*, edited by P. Liatsis, A. Hussain, S. A. Mostafa, and D. Al-Jumeily, Communications in Computer and Information Science, vol. 1548. Springer.

Segall-Corrêa, A. M., L. Marin-Leon, H. Melgar-Quiñonez, and R. Pérez-Escamilla. 2014. "Refinement of the Brazilian Household Food Insecurity Measurement Scale: Recommendation for a 14-Item EBIA." *Revista de Nutrição* 27 (2): 241–251.

Smits, N. 2010. "A Note on Youden's J and Its Cost Ratio." *BMC Medical Research Methodology* 10 (89).

Sugiyama, M., M. Krauledat, and K.-R. Müller. 2007. "Covariate Shift Adaptation by Importance Weighted Least Squares." *Journal of Machine Learning Research* 8: 985–1005.

United Nations. 2023. *The Sustainable Development Goals Report 2023: Special Edition*. https://unstats.un.org/sdgs/report/2023/.

United Nations. 2025. *The Sustainable Development Goals Report 2025*. https://unstats.un.org/sdgs/report/2025/.

UNEP. 2025. *Frontiers 2025: The Weight of Time—Facing a New Age of Challenges for People and Ecosystems.* United Nations Environment Programme. https://wedocs.unep.org/20.500.11822/48122.

Youden, W. J. 1950. "Index for Rating Diagnostic Tests." *Cancer* 3 (1): 32–35.

APPENDIX A: HYPERPARAMETER TUNING

As tuning across the full dataset or within a nested loop is computationally demanding for large national surveys, such as the PNADC, we adopted a computationally efficient "global tuning" approach. Specifically, we extracted a 20 per cent representative subsample from the globally oversampled 2023 training set, mimicking the complex survey design of the PNADC by sampling the geographic clusters (PSUs), rather than random households. It is important to note that this PSU sampling was stratified by both survey stratum and also the food security status to maintain the balanced proportion of food-insecure and food-secure households from our oversampled training set.

On the tuning subsample, we conducted an optimisation process to identify the optimal hyperparameters for each algorithm (reported in Table A1). Specifically, we tuned one hyperparameter for both the DT and the LASSO, two for the RF, three for the SVM, and six for XGBoost. The selected hyperparameters were subsequently fixed and applied consistently across all iterations of the 5-fold cross-validation, as well as in the final 2023 versus 2024 out-of-sample evaluation. Table A1 reports the final hyperparameter values selected for both specifications of each algorithm, along with the search strategy and grid range employed.

TABLE A1. Hyperparameter tuning: final selected values by model and variant

Model	Parameter	Tuning strategy and search space	w/ Precip.	w/o Precip.
DT	cp	CP xerror pruning (rpart); initial cp = 0.001	0.001	0.001
LASSO	penalty (λ)	Grid search, 10 levels; log scale [10^{-4}, 10^{0}]	0.0001	0.0001
RF	mtry	Bayesian opt., 15 iter.; finalised from training data	18	15
	min_n	Bayesian opt., 15 iter.; default dials range	40	40
XGBoost	mtry	Bayesian opt., 15 iter. + 10 initial; finalised from training data	57	27
	min_n	Bayesian opt., 15 iter. + 10 initial; default dials range	14	3
	tree_depth	Bayesian opt., 15 iter. + 10 initial; default dials range	7	6
	learn_rate	Bayesian opt., 15 iter. + 10 initial; default dials range	≈0.0499	≈0.0745
	loss_reduction	Bayesian opt., 15 iter. + 10 initial; default dials range	<0.000001	≈0.0087
	sample_size	Bayesian opt., 15 iter. + 10 initial; default dials range	0.9946	0.9159
SVM	cost	Grid search, 18 combinations; cost = log10 [-2, 2]; degree = {1,2}; scale factor = log10 [-3, -1]	1.000	1.000
	degree	Grid search, 18 combinations; cost = log10 [-2, 2]; degree = {1,2}; scale factor = log10 [-3, -1]	2	2
	scale_factor	Grid search, 18 combinations; cost = log10 [-2, 2]; degree = {1,2}; scale factor = log10 [-3, -1]	0.010	0.100

Notes: All parameters were optimised for AUROC on the 2023 training data. cp selected via xerror minimisation in rpart's internal cross-validation table. Abbreviations: DT, Decision Tree; RF, Random Forest; XGBoost, eXtreme Gradient Boosting; SVM, Support Vector Machine. w/ Precip. = model trained with precipitation variables; w/o Precip. = model trained without them; opt. = optimisation; iter. = iteration.

To maintain computational feasibility during this tuning process, we restricted the ensemble methods (RF and XGBoost) to 100 trees across the hyperparameter grid. Once the best configurations were identified, both models were retrained with 500 trees. The polynomial SVM required additional computational safeguards. Unlike tree-based methods, SVMs must internally compute a similarity score between every pair of observations in the training data, a

task that becomes substantially expensive as the sample grows. To account for this, we applied supplementary data caps specific to the SVM: the tuning subsample was limited to 5,000 households, while the training data for each CV fold and the final model were capped at 15,000 households. In all cases, this subsampling was stratified by food security status to strictly preserve the 50/50 class balance achieved through our global oversampling strategy.

Because of these data caps, the SVM was the only algorithm not exposed to the complete dataset during the training stage. However, its final evaluation on the 2024 test set was conducted on the full sample, as scoring new observations does not carry the same computational burden as model fitting. The SVM's performance results should therefore be interpreted with this constraint in mind, as it operated under materially different training conditions than its tree-based and linear counterparts.

IPCid ipea Institute for Applied Economic Research
MINISTRY OF PLANNING AND BUDGET

ipcid@ipea.gov.br ▪ www.ipea.gov.br