A Machine Learning Framework for Early Detection of Food Insecurity Using Administrative Microdata

Keila Barbosa[1,2], **Andre L. Aquino**[1]

[1]Orion Lab., Instituto de Computação
Universidade Federal de Alagoas, Maceió – AL – Brasil

[2]Centro de Informática
Universidade Federal de Pernambuco, Recife – PE – Brasil

{keilabarbosa, alla}@orion.ufal.br

Abstract. *Food insecurity is a multidimensional phenomenon strongly associated with socioeconomic, household, and territorial factors. This study proposes a machine learning pipeline to predict food insecurity risk using administrative microdata from Brazil's Cadastro Único (CadÚnico) from the state of Alagoas. The approach integrates feature engineering, predictive modeling, temporal evaluation, SHAP-based interpretability, and spatial aggregation of predictions. We evaluate both linear and tree-based models, including Logistic Regression, Random Forest, LightGBM, and CatBoost. A temporal evaluation protocol is adopted, using* 2024 *data for training/validation and* 2025 *data for testing, improving temporal generalization and evaluation realism. The final LightGBM model achieved an ROC-AUC of* 0.91 *and PR-AUC of* 0.72 *on the temporal test set, with a recall of* 0.78*, indicating strong capability in identifying at-risk households. Interpretability analysis highlights the importance of territorial and socioeconomic variables, while spatial aggregation reveals clear geographic patterns of vulnerability. The results demonstrate the potential of combining administrative data and machine learning to support early warning systems and inform data-driven public policies.*

1. Introduction

Food insecurity remains a major social challenge, particularly in developing countries, where structural inequalities directly affect access to adequate food [UNICEF et al. 2021]. In Brazil, this issue persists in a heterogeneous manner across the territory, disproportionately affecting low-income populations.

Traditional measurement approaches based solely on income are limited, as they fail to capture the multidimensional nature of food insecurity, which involves household characteristics, access to services, cost of living, and territorial factors [Sen 1982, Alkire et al. 2015]. In this context, large-scale administrative data, such as Cadastro Único (CadÚnico), enable more comprehensive analyses by providing detailed information on millions of families.

When combined with machine learning techniques, these data enable predictive approaches that identify populations at risk. Tree-based models, such as Light Gradient Boosting Machine (LightGBM), have shown strong performance in tabular data by capturing nonlinear relationships and complex interactions [Chen and Guestrin 2016,

Ke et al. 2017, Grinsztajn et al. 2022]. However, there is still a lack of studies that integrate predictive modeling, interpretability, and spatial analysis within a unified framework, particularly under temporal evaluation settings.

To address these gaps, this work proposes an integrated pipeline for predicting food insecurity risk using CadÚnico administrative microdata. The approach combines feature engineering, machine learning modeling, interpretability via Shapley Additive Explanations (SHAP), and spatial analysis, and is evaluated under a temporal protocol (training on 2024 data and testing on 2025 data).

Specifically, this study aims to answer the following research questions:

- **QP1 – Risk prediction:** Can food insecurity risk be accurately predicted using multidimensional administrative data from CadÚnico?
- **QP2 – Risk determinants:** Which household, socioeconomic, and territorial characteristics most contribute to food insecurity risk?
- **QP3 – Model comparison:** Do machine learning models outperform traditional statistical approaches in this context?

The main contributions of this work include the proposal of an integrated pipeline that combines predictive modeling, interpretability, and spatial analysis using large-scale administrative data; the adoption of a temporal evaluation protocol that provides a more realistic assessment of model generalization; empirical evidence highlighting the central role of territorial variables in predicting food insecurity; and the demonstration of the potential of machine learning-based early warning systems to support data-driven public policy design.

The results show that tree-based models, particularly LightGBM, achieve high predictive performance and strong temporal generalization, while also providing meaningful insights into the determinants and spatial distribution of food insecurity risk.

2. Related Work

Food insecurity is a multidimensional phenomenon shaped by economic, social, and territorial factors, not fully explained by income alone [Sen 1982]. Multidimensional frameworks emphasize the interplay among housing, household structure, and access to services [Alkire et al. 2015, UNICEF et al. 2021].

The growing availability of administrative data enables granular analysis of social vulnerability, particularly in public policy contexts [Einav and Levin 2014]. In parallel, tree-based machine learning models (e.g., Random Forest, XGBoost, LightGBM) have shown strong performance in tabular data by capturing nonlinear interactions [Chen and Guestrin 2016, Ke et al. 2017, Grinsztajn et al. 2022].

However, gaps remain in integrating administrative data, predictive modeling, interpretability, and spatial analysis within a realistic temporal evaluation framework. Interpretability methods such as SHAP [Lundberg and Lee 2017] and spatial techniques like LISA [Anselin 1995] are often used separately.

This work addresses these gaps by combining large-scale administrative data, machine learning, interpretability, and spatial analysis within a realistic temporal evaluation framework.

3. Methodology

The methodological workflow is illustrated in Figure 1. The process begins with raw administrative microdata, followed by data cleaning, harmonization, and transformation. Derived features are then constructed to capture structural dimensions of social vulnerability.

Next, the dataset is partitioned temporally into training/validation and test sets, simulating a real-world deployment scenario. Models are trained and evaluated, and their predictions are analyzed using interpretability techniques. Finally, predicted probabilities are aggregated at the municipal level.

This approach follows established best practices for leveraging administrative data in large-scale inference and prediction tasks [Einav and Levin 2014].

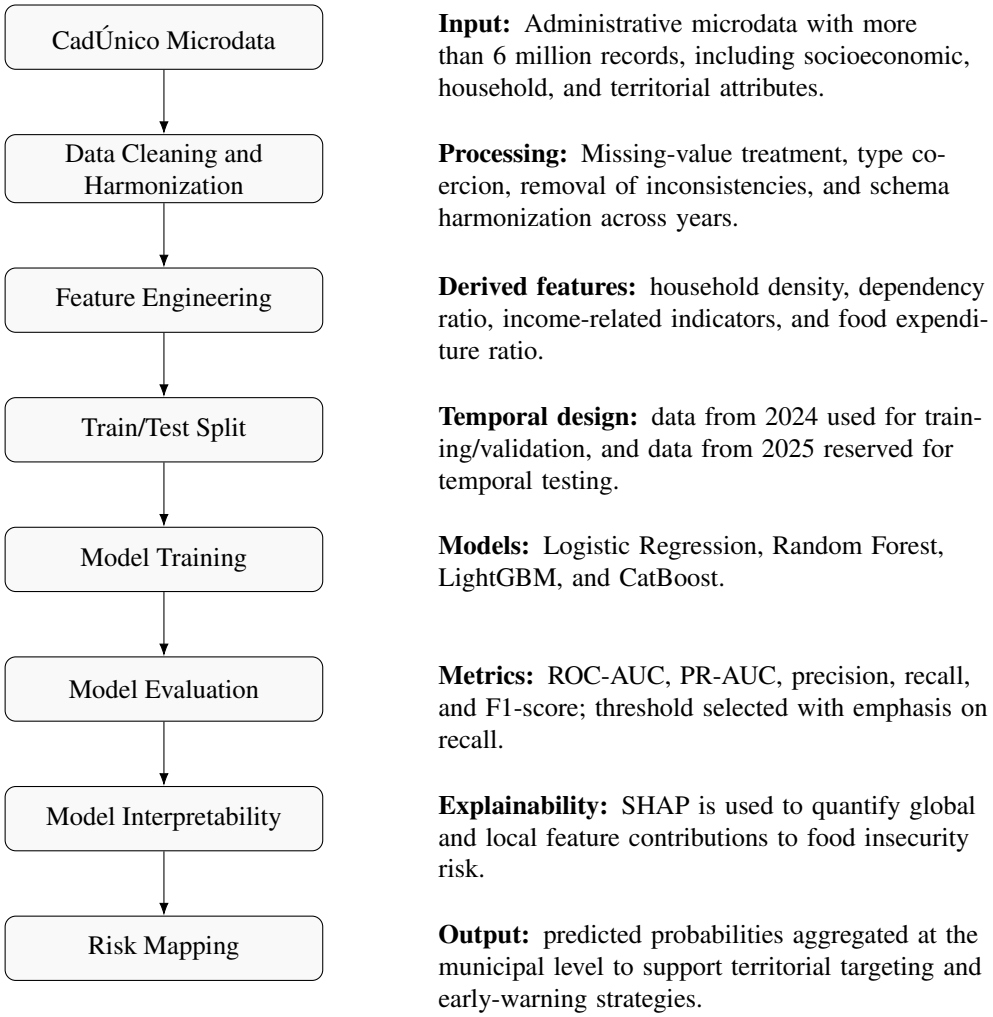

Figure 1. Detailed methodological pipeline adopted to build the food insecurity early-warning system using CadÚnico administrative microdata.

3.1. Analytical dataset construction

The dataset was built from Cadastro Único (CadÚnico) microdata for the state of Alagoas, covering the period from 2023 to 2025, with an initial volume exceeding 6 million records.

Preprocessing steps included: (i) filtering records with valid target labels, (ii) removing inconsistencies and duplicates, (iii) standardizing data types, (iv) handling missing values, (v) aligning variables across years, and (vi) removing attributes with potential data leakage.

After preprocessing, the final dataset comprised approximately 880 thousand valid records. Two versions were defined: a full dataset with 334 features and a reduced dataset with 255 features, to ensure a fair comparison of models.

Table 1 summarizes the main characteristics of the dataset, including size, dimensionality, and temporal split.

Table 1. Summary of the analytical dataset after preprocessing

Metric	Value
Total records (raw)	6,038,705
Valid records	880,359
Number of features (full dataset)	334
Number of features (modeling set)	255
Training set (2024)	449,323
Test set (2025)	431,036
Positive class prevalence	8.28%

3.2. Feature engineering

Based on the multidimensional poverty framework [Alkire et al. 2015], feature engineering was performed to capture structural aspects of social vulnerability not directly represented in the original variables.

Derived features were created to reflect household composition, economic pressure, and living conditions, improving the representation of the problem and enhancing predictive performance.

The main transformations include:

Household density:

$$density = \frac{qtd\_pessoas\_domic\_fam}{qtd\_comodos\_dormitorio\_fam}$$

A proxy for overcrowding, often associated with poor housing conditions.

Dependency ratio:

$$dependents = \frac{qtd\_pessoas\_0 - 17 + qtd\_pessoas\_65+}{qtd\_pessoas\_domic\_fam}$$

Measures the proportion of economically dependent individuals in the household.

Food expenditure ratio:

$$food\_ratio = \frac{val\_desp\_alimentacao\_fam}{vlr\_renda\_total\_fam}$$

Represents the share of income spent on food, indicating economic vulnerability.

In addition, key original variables were retained, including territorial indicators (e.g., IBGE code), socioeconomic variables (income and expenses), household characteristics, and temporal attributes, enabling the model to capture both structural and contextual factors related to food insecurity.

3.3. Predictive modeling and Evaluation metrics

We evaluated multiple machine learning models, including Logistic Regression, Random Forest, LightGBM, and CatBoost (Table 2), enabling comparison between linear and nonlinear approaches.

Table 2. Models evaluated

Model	Description
Logistic Regression	Linear baseline model
Random Forest	Tree-based ensemble
LightGBM	Histogram-based gradient boosting
CatBoost	Gradient boosting with categorical support

Tree-based models were emphasized due to their ability to capture nonlinear relationships and feature interactions in tabular socioeconomic data [Grinsztajn et al. 2022]. All models were trained exclusively on 2024 data, following the temporal protocol, and evaluated on the 2025 test set to ensure robustness.

LightGBM was selected as the final model based on its superior validation performance compared to Random Forest and CatBoost.

To improve temporal generalization and avoid data leakage, a temporal split was adopted, separating observations from different periods.

Data from 2024 were used for training and validation, while 2025 data were reserved for testing, simulating a real-world prospective scenario.

Although temporal dependencies are not explicitly modeled, the temporal split simulates a realistic deployment scenario.

Table 3 summarizes the dataset distribution. The class prevalence remains stable across periods, supporting a reliable assessment of model generalization.

Table 3. Dataset distribution under the temporal protocol

Set	Instances	Features	Positives	Prevalence
Train/validation (2024)	449,323	334	36,831	0.08197
Test (2025)	431,036	334	36,039	0.08361

Due to class imbalance (approximately 8% positive cases), robust evaluation metrics were adopted to properly assess model performance in rare-event settings [He and Garcia 2009, Haixiang et al. 2017].

Table 4 summarizes the metrics used.

Table 4. Evaluation metrics

Metric	Formula	Description
Precision	$\frac{TP}{TP+FP}$	Proportion of predicted positives that are correct
Recall	$\frac{TP}{TP+FN}$	Ability to identify true positive cases
F1-score	$2 \cdot \frac{P \cdot R}{P+R}$	Balance between precision and recall
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness (sensitive to imbalance)
ROC-AUC	$\int_0^1 TPR(FPR)\, d(FPR)$	Overall class separability across thresholds
PR-AUC	$\int_0^1 Precision(Recall)\, d(Recall)$	Performance focused on the positive class

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives.

Recall was prioritized as the main evaluation criterion to minimize false negatives, which is critical in identifying vulnerable households. PR-AUC was used as a complementary key metric due to its higher sensitivity in imbalanced scenarios [Saito and Rehmsmeier 2015].

Although ROC-AUC is widely used, it may yield overly optimistic results in imbalanced settings, as it is influenced by the majority class. In contrast, PR-AUC focuses on the trade-off between precision and recall, providing a more informative assessment of rare-event detection [Saito and Rehmsmeier 2015, He and Garcia 2009].

Given the application context, threshold selection was guided by recall-oriented metrics to ensure accurate alignment and identify at-risk populations.

3.4. Interpretability, robustness analysis, and spatial aggregation

After model training and evaluation, additional analyses were conducted to interpret model behavior, assess robustness, and examine results at the territorial level.

Interpretability was performed using Shapley Additive Explanations (SHAP) [Lundberg and Lee 2017], enabling both global and local analysis of feature contributions and providing insights into the main drivers of food insecurity risk.

Model robustness was evaluated through complementary experiments, including an ablation study removing territorial variables, probability calibration to improve prediction reliability, and hyperparameter optimization using Optuna.

Finally, predicted probabilities were aggregated at the municipal level as:

$$Risk_m = \frac{1}{N_m} \sum_{i=1}^{N_m} P_i$$

where P_i is the predicted probability for household i and N_m is the number of

households in municipality m.

This aggregation produces a continuous territorial risk indicator, which was linked to municipal geographic data (IBGE code) to generate choropleth maps. This enables the identification of spatial patterns and regional disparities in food insecurity, supporting evidence-based public policy. Calibration and tuning experiments were used to assess model robustness and performance sensitivity, and hyperparameter optimization informed the final model selection.

4. Results

This section presents the empirical evaluation of the proposed pipeline in four stages: (i) baseline model comparison, (ii) methodological assessment (ablation, calibration, and hyperparameter tuning), (iii) final model selection and temporal generalization, and (iv) interpretability and spatial analysis of predicted risk.

Models were first evaluated on the validation set to identify the most competitive approach. Subsequent experiments assessed the impact of removing territorial variables, probability calibration, and hyperparameter optimization. The selected model was then retrained using the full 2024 dataset and evaluated on the 2025 temporal test set.

4.1. Baseline and Initial Model Comparison

Table 5 reports the performance of baseline and initial models on the validation set. Ran-

Table 5. Baseline and initial model comparison on the validation set

Model	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Logistic Regression	0.08	0.08	1.00	0.15	0.485	0.08
Random Forest	**0.94**	**0.60**	**0.74**	**0.66**	**0.93**	**0.77**
LightGBM (weighted)	0.71	0.18	0.70	0.29	0.78	0.26
LightGBM (unweighted)	0.71	0.18	0.69	0.28	0.78	0.31
CatBoost	0.71	0.17	0.67	0.27	0.76	0.23

dom Forest achieves the best performance among the baseline models, serving as a strong initial benchmark. In contrast, Logistic Regression exhibits a degenerate behavior, achieving maximum recall at the expense of extremely low precision and accuracy, rendering it impractical.

Although Random Forest achieves a slightly higher F1-score in the baseline setting, LightGBM was selected for its greater flexibility and ability to achieve substantial performance improvements after hyperparameter optimization.

Figure 2 illustrates the ROC and Precision-Recall curves for the evaluated baseline models.

4.2. Methodological Experiments on the Validation Set

Following the initial comparison, a focused methodological analysis was conducted on LightGBM under four scenarios: baseline validation model, removal of territorial variables, probability calibration, and hyperparameter tuning via Optuna.

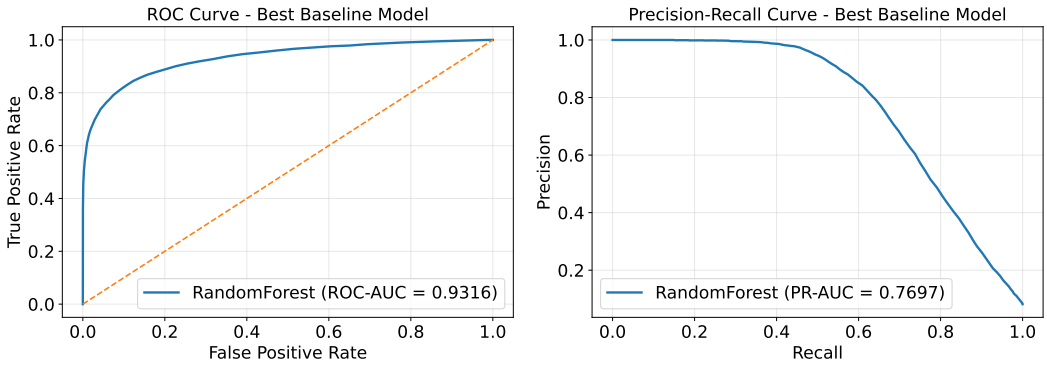

Figure 2. ROC and Precision–Recall curves for baseline models.

Table 6. Methodological experiments on the validation set

Experiment	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
Base (validation)	0.83	0.23	0.47	0.31	0.77	0.25
Without territorial variables	0.86	0.25	0.37	0.30	0.75	0.26
Calibrated	0.84	0.24	0.44	0.31	0.77	0.25
Tuned (Optuna)	**0.95**	**0.68**	**0.63**	**0.65**	**0.93**	**0.72**

Table 6 summarizes the results. Removing territorial variables leads to a clear degradation in performance, particularly in recall and ROC-AUC, indicating that spatial context carries substantial predictive information. Probability calibration yields negligible improvements over the baseline.

In contrast, hyperparameter tuning produces substantial gains across all metrics, highlighting LightGBM's sensitivity to proper configuration. However, these results should be interpreted cautiously, as optimization was performed on the validation set, potentially introducing optimistic bias.

Future work should include systematic hyperparameter optimization for all models to ensure a fully controlled comparison.

4.3. Final Model and Temporal Generalization

Based on the methodological experiments, the tuned LightGBM was selected as the final model.

A threshold of 0.09 was selected based on the precision–recall trade-off, prioritizing high recall to minimize false negatives, a critical goal for early warning systems.

The final model was configured with a learning rate of 0.05, 63 leaves, a minimum of 20 samples per leaf, subsampling and feature sampling rates of 0.8, no $L1$ regularization, $L2$ regularization of 1.0, and 2000 boosting iterations.

The model was retrained on the full 2024 dataset and evaluated on the 2025 temporal test set. Table 7 reports its performance.

On the temporal test set, the model maintains strong discriminative performance (ROC-AUC = 0.91, PR-AUC = 0.72) and high recall (0.78), confirming its effectiveness

Table 7. Final model performance on the 2025 temporal test set

Model	Threshold	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC
LightGBM (final)	0.09	0.88	0.40	0.78	0.53	0.91	0.72

in identifying at-risk households. The lower precision (0.40) reflects an expected trade-off driven by the recall-oriented threshold, increasing coverage of positive cases at the cost of more false positives.

Figure 3 presents the ROC and Precision-Recall curves for the temporal test set.

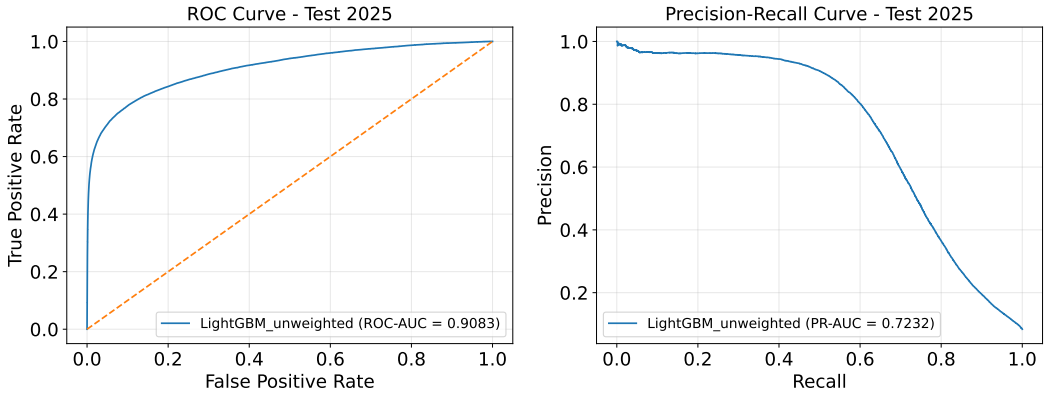

Figure 3. ROC and Precision–Recall curves on the 2025 temporal test set.

4.4. Model Interpretability

Figure 4 shows the global feature importance estimated via SHAP, while Table 8 summarizes the contribution by feature group.

The interpretability analysis reveals that territorial variables dominate the model's predictions, followed by socioeconomic and temporal factors. This pattern is consistent with the ablation results, reinforcing the central role of spatial context in predicting food insecurity risk.

Table 8. Relative contribution by feature group

Group	Mean SHAP	Percentage (%)
Territorial	1.463	45.82
Socioeconomic	0.711	22.27
Temporal	0.559	17.52
Other	0.460	14.39

4.5. Spatial Distribution of Predicted Risk

Predicted food insecurity risk corresponds to the household-level probability of vulnerability estimated from socioeconomic, administrative, and territorial variables in

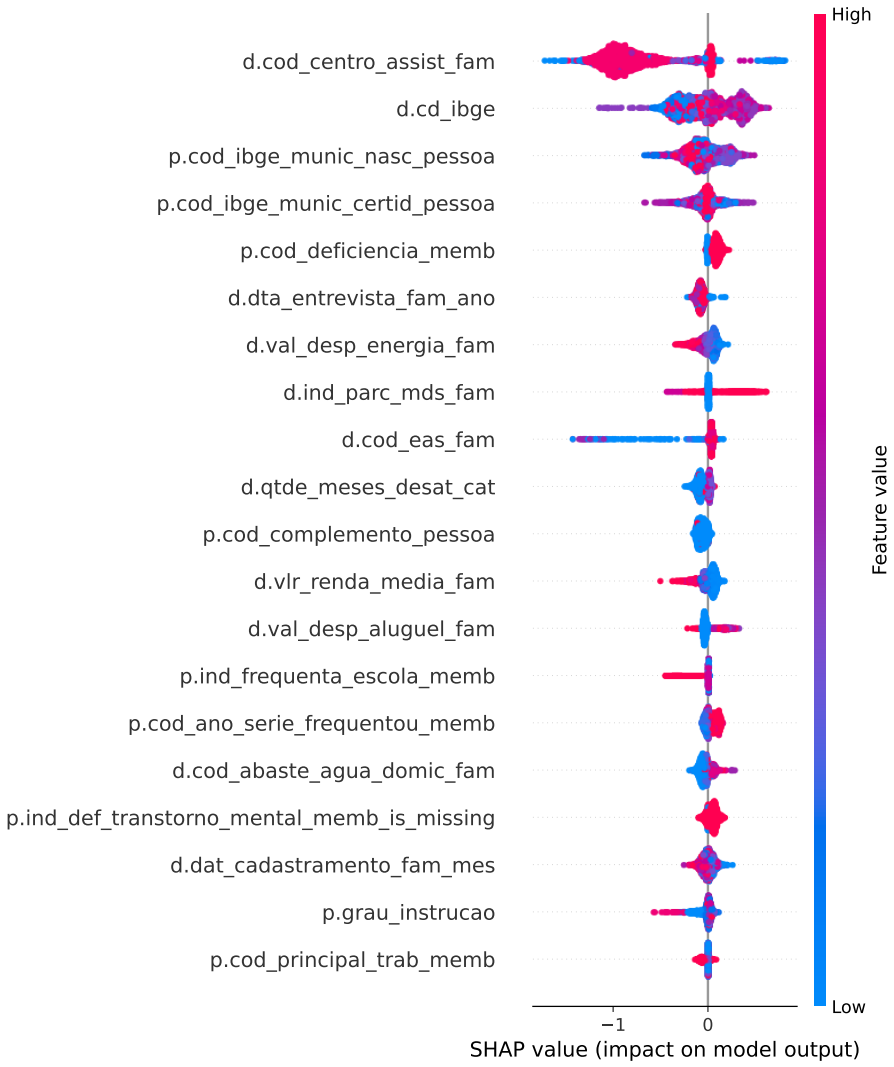

Figure 4. Global feature importance based on SHAP values.

CadÚnico. For spatial analysis, these probabilities were aggregated to the municipal level using the final model's mean prediction.

Figure 5 shows the spatial distribution of predicted risk across Alagoas in 2025, presented both as a continuous surface and by quantiles. Table 9 lists the municipalities with the highest average predicted risk. The maps reveal pronounced territorial heterogeneity, with higher risk levels concentrated in specific municipalities. Combined with the SHAP and ablation results, this pattern indicates that spatial factors are not merely descriptive but structurally embedded in the model's predictive capacity.

Overall, the results confirm that food insecurity is a multidimensional phenomenon driven by the interaction of socioeconomic, structural, and territorial factors, and that administrative data can effectively support large-scale predictive systems. Tree-based models achieved strong performance (ROC-AUC ≈ 0.93; PR-AUC ≈ 0.77) and maintained robust generalization in the 2025 temporal test (ROC-AUC ≈ 0.91; PR-AUC ≈ 0.72), indicating stable predictive relationships over time.

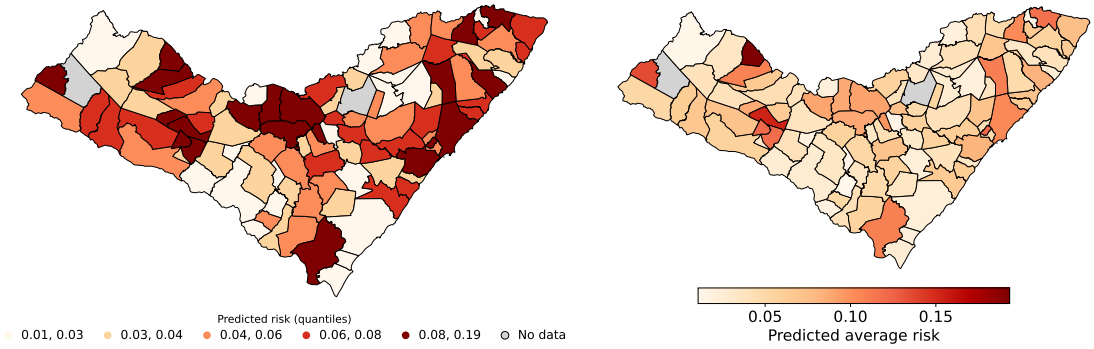

Figure 5. Spatial distribution of predicted risk in 2025: quantile-based map (left) and continuous scale (right).

Table 9. Municipalities with the highest predicted risk in 2025

IBGE Code	Mean Risk	Observed Prevalence	Number of Households
2706109	0.193	0.247	4509
2705705	0.154	0.223	2768
2706422	0.136	0.211	1865
2707909	0.123	0.163	851
2705408	0.122	0.157	2134

Territorial variables emerge as the main drivers of risk, reinforcing that vulnerability is strongly context-dependent. This is further supported by ablation results showing performance degradation when spatial features are removed. Tree-based models consistently outperform linear approaches, with LightGBM providing the best balance after tuning. The recall-oriented threshold reflects a deliberate trade-off in favor of early detection, supporting the model's use in territorialized, proactive policy interventions.

Although the spatial analysis is descriptive, it reveals clear territorial patterns that suggest spatial dependence, motivating future work using formal spatial statistical methods.

5. Conclusion

This work proposes a unified pipeline for predicting food insecurity risk using administrative data, integrating machine learning, temporal evaluation, interpretability, and spatial analysis.

Results show that tree-based models, particularly LightGBM, achieve strong predictive performance and temporal generalization. Territorial and socioeconomic factors emerge as the main drivers, and spatial aggregation reveals clear geographic concentration patterns.

The proposed approach demonstrates the practical feasibility and policy relevance of large-scale, data-driven early warning systems.

A limitation of this study is the absence of external validation using independent food insecurity indicators, which should be addressed in future work.

The computational artifacts are publicly available in Repository.[1]

5.1. Future Work

Future work includes: (i) dynamic and longitudinal modeling for change detection, (ii) integration of external data sources (e.g., prices, climate), (iii) formal spatial econometric modeling, (iv) external validation with established indicators, and (v) deployment and evaluation in real policy environments.

Acknowledgements

Research Support Foundation of the State of Alagoas (FAPEAL), through grant E:60030.0000000352/2021, and by the National Council for Scientific and Technological Development (CNPq), through grant 407515/2022-4.

References

Alkire, S., Roche, J. M., Ballon, P., Foster, J., Santos, M. E., and Seth, S. (2015). *Multidimensional poverty measurement and analysis*. Oxford University Press, USA.

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical Analysis*, 27(2):93–115.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Einav, L. and Levin, J. (2014). Economics in the age of big data. *Science*, 346(6210):1243089.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 73:220–239.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432.

Sen, A. (1982). *Poverty and famines: An essay on entitlement and deprivation*. Oxford university press.

UNICEF et al. (2021). *The state of food security and nutrition in the world 2021*. FAO;.

[1] https://github.com/keilabcs/CadUnicoIA