






Machine Learning methods and models to predict food insecurity levels for families in Ceará, Brazil, based on employment, housing and other social indicators



Ticiania L. Coelho da Silva   [Insight Data Science Lab, Federal University of Ceará, Brazil | ticianalc@insightlab.ufc.br]


Lara Sucupira Furtado  [Transportation Engineering, Federal University of Ceará, Brazil | lfurtado@det.ufc.br]


Guilherme Sales Fernandes   [Insight Data Science Lab, Federal University of Ceará, Brazil | guilherme.sales@insightlab.ufc.br]

José A. Fernandes de Macêdo   [Insight Data Science Lab, Federal University of Ceará, Brazil | jose.macedo@insightlab.ufc.br]

Livia Almada Cruz   [Insight Data Science Lab, Federal University of Ceará, Brazil | livia@insightlab.ufc.br]

Regis Pires Magalhães   [Insight Data Science Lab, Federal University of Ceará, Brazil | regis@insightlab.ufc.br]

Laécia Gretha Amorim Gomes  [Social Protection Secretariat, Ceara State Government | laeciagomes@gmail.com]

 *Insight Data Science Lab, Federal University of Ceará, Brazil.*

Received: 10 January 2023 • **Accepted:** 13 January 2025 • **Published:** 25 March 2025

Abstract Many nations still struggle to provide their populations with access to food and balanced nutrition. The Food and Agriculture Organization of the United Nations (FAO) included Brazil in its 2022 Hunger Map, highlighting that 61 million Brazilians face difficulties in feeding themselves. Despite the presence of various food security alert and monitoring systems in food-insecure countries, the data and methodologies they rely on capture only a fraction of the issue's complexity, underscoring the need for further research to fully comprehend this multifaceted problem. In response, the Secretary for Social Protection of Ceará (SPS - Secretaria de Proteção Social), located in Brazil's northeast, conducted a survey to collect data on the social and economic characteristics of extremely vulnerable families. This dataset, analyzed in our study, represents a concentrated effort by the government of Ceará to evaluate the needs of low-income households, particularly those with children who lack access to essential services. We used the Brazilian Food Insecurity Scale, a tool validated by the Brazilian Ministry, to measure food insecurity levels based on families' responses, assigning scores to their answers. This paper presents a machine learning model that examines the collected data to identify which factors related to Food Access, Employment and Income, Housing, and Public Services can predict levels of food insecurity. Our best model demonstrates an accuracy of approximately 0.75, an F1-score of 0.80, and can distinguish between severe and non-severe food insecurity levels. We suggest that our model could be applied to other datasets lacking nutrition-specific questions to gauge a family's food insecurity level. Additionally, our research sheds light on the key factors influencing food insecurity levels in Brazil, notably income and housing conditions, providing valuable insights for addressing this issue.

Keywords: food insecurity, machine learning, feature importance, classification model

1 Introduction

Addressing social challenges often requires comprehensive data collection to accurately understand issues and design effective public policies [Furtado and Furtado, 2021]. Food insecurity, a global concern, exemplifies this need. Defined as the struggle to access regular, nutritious, and affordable food [Costa *et al.*, 2017], food insecurity affects an estimated 36% of Brazilian families — a situation worsened by the COVID-19 pandemic [Neri, 2022]. Factors such as household income, the presence of children, and housing instability are closely linked to food insecurity [Lee *et al.*, 2021], underscoring its complexity and the importance of continuous monitoring at both national and local levels. However,

in-person surveys to gather this data are often limited by high costs, significant time demands, and the challenges of reaching extremely vulnerable families [Nica-Avram *et al.*, 2020].

Despite these challenges, the Secretary for Social Protection of Ceará (SPS - Secretaria de Proteção Social) in northeast Brazil undertook a major effort to collect data on the social and economic characteristics of highly vulnerable families. By August 2022, the CMIC (*Cartão Mais Infância Ceará*) Survey had reached 184 municipalities, gathering data from nearly 50,000 families. CMIC is a monthly cash transfer program from the Ceará State Government aimed at supporting extremely poor families. This paper analyzes that dataset, which represents a focused government effort in Ceará to assess the needs of low-income households, particu-

larly those with children lacking access to essential services. This systematic approach highlights Ceará's commitment to addressing food insecurity by exploring its underlying factors within the state's context.

The comprehensive dataset contains 183 questions, covering topics from children's development to the health of parental figures, including the mother's physical and mental well-being. For this study, we selected 29 questions focused on the family unit at the household level, providing a broader overview of the situation.

A crucial component of the CMIC dataset includes 5 questions on food insecurity and 3 questions on food access. These questions explore the frequency with which individuals are unable to purchase food and their experiences of hunger. The dataset uses the Brazilian Food Insecurity Scale (EBIA, *Escala Brasileira de Insegurança Alimentar*), a tool endorsed by the Brazilian Ministry, to quantify responses. Each question is scored, enabling the calculation of a family's level of food insecurity. This approach is unique to the state of Ceará, as no other Brazilian state has conducted a survey that integrates detailed nutrition data with broader social variables. Typically, other surveys limit their application of EBIA questions to the municipal level [Segall-Corrêa and Marin-Leon, 2009], without capturing the wider socioeconomic context.

Recognizing the urgent need for policies that address food insecurity and the unique nature of our dataset within the Brazilian context, this study aims to identify predictive features of food insecurity. We develop a machine learning model that utilizes data on Employment and Income, Housing, and Public Services to forecast food insecurity levels. While the machine learning approach itself is not groundbreaking, its application here demonstrates substantial real-world relevance, with potential to streamline data collection and guide policy development. We also propose that our model could be adapted for use with other datasets lacking nutrition-specific questions to assess food insecurity levels. This research is guided by the following key questions:

- **(RQ1)** Can a machine learning model accurately determine a family's food insecurity level using indicators related to Employment and Income, Housing, and Public Services?
- **(RQ2)** What is the simplest version of this model, in terms of reduced features, that still maintains predictive accuracy?
- **(RQ3)** Can the model's accuracy in classifying food insecurity be improved by reducing the number of features in the dataset and simplifying the classification categories?

Our proposed approach not only underscores the potential for impactful analysis but also lays the groundwork for more effective data collection methods and policy development, ultimately aiming to mitigate food insecurity through informed interventions. The remainder of this paper is organized as follows: Section 4 reviews related work; Section 5 describes the data source and transformation process; Section 5 provides an overview of our methodology; Section 6 presents the experimental evaluation; and finally, Section 7 concludes the

paper, summarizing key findings and discussing their implications.

2 Background

This section explores the theoretical foundations of the study, starting with a detailed exploration of food insecurity—specifically its definition in the context of vulnerability in Brazil—and then outlining the key concepts that frame this work.

Food Insecurity

Food insecurity involves nutritional deficits resulting from missed meals and reduced food quantity and quality [Reis, 2012]. More broadly, it refers to inadequate or uncertain access to nutritionally sufficient and safe foods, or the inability to acquire acceptable foods through socially acceptable means [Felker-Kantor and Wood, 2012]. Food insecurity is notably more prevalent in impoverished households compared to wealthier ones. However, the structure of a household also plays a role in its vulnerability to food insecurity, with female-headed households and those with children being more likely to experience moderate to severe food insecurity. Furthermore, the risk of food insecurity increases with housing instability, such as the inability to consistently pay rent [Lee et al., 2021].

In developed countries like the United States, nutritional issues often arise from poor-quality diets [Reis, 2012]. In contrast, in developing countries such as Brazil, food insecurity is assessed through a series of questions designed to measure food shortages. These questions are part of the EBIA scale and include inquiries such as, “In the last 3 months, have members of this household worried that food would run out before they could buy or receive more?” and “In the last 3 months, has anyone in the household gone hungry but did not eat because there was no money for food?” [Felker-Kantor and Wood, 2012]. This method of evaluating food insecurity underscores that vulnerability is context-specific, defined by a lack of access to basic food necessities [Reis, 2012].

Feature Selection

Feature selection is a technique used to identify a subset of relevant features from the original set by removing irrelevant, redundant, or noisy attributes. This process typically leads to better model performance, including higher accuracy, reduced computational costs, and improved interpretability [Sammut and Webb, 2017]. In our study, we applied feature selection to identify the most impactful attributes for classifying food insecurity.

To determine the optimal subset of features, we explored several strategies, including methods such as the Chi-square test, which assesses the dependency between stochastic variables and identifies features most likely to be independent of the target class. Independent or less relevant features were excluded from the model. This approach was combined with p-value analysis, where a low p-value provides strong

evidence against the null hypothesis (e.g., feature independence), and feature variance evaluation, which removes features with very low variance due to their limited discriminatory power and minimal contribution to the model's performance. Further details on the application of these methods in this paper can be found in Section 5

Feature selection methods for classification tasks are typically categorized into three main types: filter models, wrapper models, and embedded models [Tang *et al.*, 2014]. Filter approaches focus primarily on the intrinsic properties of features within the training data. In contrast, wrapper methods evaluate the quality of features based on classifier performance. Embedded methods combine elements of both filter and wrapper approaches; they assess various feature subsets using statistical criteria and select the subset that maximizes accuracy.

The feature selection methods applied in training the classification models for this study includes Select From Model (SFM), an embedded method that selects a subset of features based on their importance scores provided by a classifier. Additionally, we implemented Recursive Feature Elimination (RFE), another embedded approach that trains a classifier with all features and then recursively removes the least important features until a predefined number of features remains. We also applied Sequential Feature Selector (SFS), a greedy algorithm that decides whether to include or exclude features based on a classifier's cross-validation score. The wrapper method we employed is Boruta, which uses a random forest algorithm. Boruta evaluates feature relevance by comparing the importance of actual features with that of randomly generated probes. Finally, we used Select K Best to select features based on the highest scores from Chi-squared tests, focusing on the top k scoring features for our experiments.

The use of feature selection methods alongside the training of classification models is not a mandatory step. It is worth noting that in all cases, if the optimal model performs best with all attributes, they will be retained without modification.

Automatic Machine Learning and Classification Models

Automatic machine learning (AutoML) is a widely used and promising solution for building AI systems without human assistance to overcome challenges when it comes to creating more general, flexible, and free-of-human bias models and identifying search spaces [He *et al.*, 2021]. The success of ML models depends on the proper selection of algorithms and hyperparameters. This setup is difficult and, as a result, models often do not reach their full potential. To facilitate configuration, automatic machine learning (AutoML) systems focus mainly on hyperparameter optimization to achieve the best possible prediction models.

AutoML experiments different ensemble-based algorithms, which typically exhibit high accuracy [Caruana and Niculescu-Mizil, 2006]. They have recently been successfully applied to a large number of prediction and classification issues [Zhou, 2012]. Tree boosting [Schapire *et al.*, 1998] and bagging [Breiman, 1996] are two frequently used ensemble approaches. They are broad strategies that can im-

prove precision of the predictions made by tree-based systems. Several tree predictors trained on bootstrap samples of the training data are combined in tree bagging, also known as bootstrap aggregation. To generate the overall forecast, the data is pooled using simple averaging for regression and simple voting for classification, with the variance being reduced due to the averaging.

According to Sutton [2005], there are some differences between bagging and boosting: In contrast to bagging, which employs a straightforward average of results to provide an overall forecast, boosting makes use of a weighted average of outcomes from the application of a prediction method to various samples. Additionally, with boosting, the samples that are utilized at each step are not all selected uniformly from the same population, but rather, the cases that were incorrectly predicted in a previous phase are given more weight in the subsequent step. As a result, unlike bagging, which is based solely on an average of predictions, boosting employs weights throughout its iterative process. Additionally, boosting is frequently used with weak learners (e.g., a simple classifier such as a two-node decision tree), although this is not the case with strong learners.

Random forests (RF) proposed by Breiman [2001] add an additional layer of randomness to bagging. For the collection of splitting variables, it employs random feature selection at each node. RF are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. As the number of trees in a forest increases, the generalization error converges to a limit. The strength of the individual trees in the forest and their association with one another determine the generalization error of the forest. Each node is split using a random sample of features, which results in lower error rates and greater noise resistance. Internal estimates keep track of error, strength, and correlation; they are used to demonstrate how the splitting process responds to an increase in the number of features. Internal estimations are another method for evaluating variable significance.

Gradient Boosting [Friedman, 2001] is another boosting technique with the primary goal of creating a chain of weak models; each one aims to minimize the error of the previous model through a loss function. Each weak model (a decision tree, generally) is adjusted based on the learning rate. The algorithm works as follows initially; a model is rated. We compute an error from the predicted to the real value. Another model is created and adjusted based on the error from the previous model. The process is repeated several times to minimize the error between the weak models. The final model is the sum of the adjustments performed on the weak learners.

A typical neural network is a class of machine learning approach consisting of numerous interconnected neurons that produce a series of real-valued activations. The neurons are activated by the external inputs or through weighted connections from previous neurons [Schmidhuber, 2015]. The primary goal of its learning algorithm is to determine the best weight configuration for a given task. Sometimes, the problem requires a long chain of stages that usually employ nonlinear transformations. This is where deep learning comes into play, which refers to neural networks that involve learn-

ing across multiple stages.

Handling Imbalanced Datasets for Classification Models

An imbalanced classification problem occurs when the distribution of samples across the known classes is biased or skewed. This imbalance can range from a slight bias to a more severe disparity. In other words, the number of samples in the minority class is smaller than in the majority class, often by a significant proportion (e.g., twice as large or more) [Fernández *et al.*, 2008].

Most classifiers assume that the data across classes is balanced and evenly distributed. However, data imbalance can lead to unexpected errors and even serious consequences in data analysis, particularly in classification tasks, as the algorithm may fail to learn patterns in the minority classes. Our food insecurity dataset is imbalanced, a challenge that we address in this paper.

Generally, there are two main strategies for addressing the imbalanced dataset problem: re-sampling and cost-sensitive re-weighting [Haixiang *et al.*, 2017]. Re-sampling approaches modify the dataset by adding or removing samples, using one of three methods:

1. Over-sampling: increases the number of samples in the minor class to balance the classes (main techniques are SMOTE and randomly duplicating the minor samples);
2. Under-sampling: eliminates samples from the major class (an effective method is random sampling);
3. Hybrid methods: combines the methods of oversampling and under-sampling.

Cost-sensitive re-weighting approaches adjust the loss function by assigning higher costs to examples from the minority class and lower costs to those from the majority class. This allows us to control the level of imbalance in the dataset. In this paper, we apply SMOTE-Tomek, a hybrid method proposed by Batista *et al.* [2003], which combines the oversampling strategy SMOTE [Chawla *et al.*, 2002] with the under-sampling strategy Tomek Link [Tomek, 1976]. SMOTE is a widely used oversampling technique that generates synthetic samples rather than resampling with replacement. For each example in the minority class, the SMOTE algorithm creates a synthetic point based on its k nearest neighbors. The concept of Tomek Links, introduced by Tomek [1976], refers to pairs of samples where one may contain noise or both are borderline. In the SMOTE-Tomek approach, synthetic points are created using SMOTE, while the undersampling strategy removes any points that form a Tomek Link. This method can remove samples from either the minority class—since synthetic examples were generated for it—or the majority class. The SMOTE-Tomek method helps reduce overfitting caused by exact replication of minority class examples, while also avoiding the loss of valuable data.

Model evaluation metrics

This section details the evaluation metrics to assess the performance of our solution.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy indicates the percentage of correct predictions, the sum of true positive and true negative divided by all categories.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Precision is the fraction between the relevant instances among the retrieved instances, the true positive values divided by the sum of true positive and false positive.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Recall is the fraction of relevant instances that were retrieved, the true positive values divided by the sum of true positive and false negative.

$$F1-Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

F1-score is the harmonic mean of the precision and recall, two times the true positive value divided by 2 times the sum of true positive, false positive and false negative values.

Feature importance

The widespread use of AI models has raised questions and concerns about their interpretability within both research and business communities [Hanif *et al.*, 2023]. The lack of interpretability in many systems' decision-making processes represents a significant shortcoming that can have serious consequences. Saarela and Jauhiainen [2021] highlight the importance of explainable artificial intelligence (XAI) in helping users and developers understand the behavior of machine learning models. Feature importance is a key technique in XAI, used to assess the relevance or contribution of each feature in predicting the target variable or outcome. There are both global and local approaches to measuring feature importance. Global methods provide an overview of feature relevance across the entire dataset, while local methods focus on understanding feature importance for individual instances or predictions. Various techniques exist for computing feature importance, including statistical correlation scores, coefficients from linear models, decision tree-based methods, permutation importance scores, SHapley Additive exPlanations (SHAP), and others [Guidotti *et al.*, 2018].

In linear models such as Linear Regression or Logistic Regression, the coefficients associated with each feature indicate its importance. Larger coefficients typically suggest higher importance.

In decision tree-based algorithms like Random Forests and Gradient Boosting Machines (GBMs), feature importance is determined by how much each feature reduces impurity across all the trees in the ensemble. This reduction in impurity is often measured using metrics like Gini impurity or entropy.

For neural networks, including deep learning models, the H2O AutoML tool uses the Gedeon method [Gedeon, 1997]

to compute feature importance. This method considers the weights connecting the input features to the hidden layers.

Permutation importance is a model-agnostic technique that assesses feature importance by randomly shuffling the values of each feature and observing the impact on the model's performance. A significant drop in performance after shuffling a feature indicates that the feature is important.

SHAP is a model-agnostic method for explaining individual predictions [Lundberg and Lee, 2017]. It is based on Shapley values, a concept borrowed from cooperative game theory, which provide a fair way to distribute the "payout" or contribution among the features in a predictive model. The primary goal of SHAP is to explain the prediction of an instance x by calculating the contribution of each feature to that prediction. In this context, feature values act as players in a coalition, and Shapley values determine how to fairly allocate the prediction among the features [Molnar, 2022]. SHAP feature importance is quantified as the mean absolute Shapley values, with features having large absolute Shapley values considered more important than those with lower values. SHAP values maintain the core principles of Shapley values while adapting them to the context of feature importance in predictions.

3 Research Problem

This section describes the research problem considered in this work.

Let VF be a set of n vulnerable families and F the set of m features related to welfare, employment, income, and housing. Let $D_F \in \mathbb{R}^{n \times m}$ be the dataset that depicts the information of each family $o \in VF$ according to all features $f \in F$. Let Y be the food insecurity level for each family $o \in VF$. Our research problem consists in finding a classification model that can learn from D_F and Y to classify vulnerable families into a food insecurity level. We have used the four metrics described in Section 2 to rank the best models execution: Accuracy (1), Precision (2), Recall (3) and F1-Score (4).

4 Related Studies

This section reviews related researches on the prediction of food insecurity, highlighting the variety of methodologies used.

Given the complex relationship between food insecurity and various social factors, this study leverages computational advancements to develop a predictive model for food insecurity, drawing insights from relevant research.

For instance, research by Lentz *et al.* [2019] utilizes linear regression to predict three measures of food insecurity: the reduced Coping Strategies Index (rCSI), the Household Dietary Diversity Score (HDDS), and the Food Consumption Score (FCS). This study combines data from multiple sources, including Integrated Food Security Phase Classification (IPC) assessments, environmental and geographic variables such as precipitation, market prices, soil quality, and remotely sensed data, as well as demographic and financial

data from household surveys. Notably, the analysis also incorporates data from mobile phone companies and remote sensing, such as the percentage of households with cell phone ownership, across various spatial scales. The dataset in Lentz *et al.* [2019] includes two surveys, one covering approximately 12,000 families and the other around 4,000, showcasing the scalability and diversity of data sources used to inform food insecurity predictions.

The study by Christensen *et al.* [2021] also focuses on predicting food crises by analyzing a World Bank dataset that includes 21 features. These features help assess the likelihood of food crises occurring in various countries from 2007 to 2020, based on criteria defined by the Famine Early Warning Systems Network (FEWS NET) and the Integrated Food Security Phase Classification (IPC) system. Their methodology employs logistic regression and classical feed-forward neural networks, with various feature selection techniques to enhance model performance. Among these approaches, the neural network model outperformed the others, achieving an F1-score of 0.84 on the test set for predicting food crises within the same year.

Similarly, a study using the 2008 Afghanistan National Risk and Vulnerability Assessment (NRVA) Survey employed decision trees and random forest models to identify household characteristics associated with food access in Afghanistan [Gao *et al.*, 2020]. This research estimated food security based on food consumption quantities and the required per capita calorie intake, focusing on factors such as household size, income, access to resources, assets, and farm production. In another application of advanced analytical techniques, [Deléglise *et al.*, 2020] utilized deep learning and machine learning models to predict food consumption and diversity indicators, drawing on data from the Burkina Faso government's permanent agricultural survey, which has been available since 2009. The survey used in Gao *et al.* [2020] sampled approximately 20,000 households across Afghanistan, illustrating the large-scale data applied in understanding food security challenges.

Another notable contribution to the field is made by Deléglise *et al.* [2022], who introduces machine learning and deep learning models to estimate the Food Consumption Score (FCS) and Household Dietary Diversity Score (HDDS). These models utilize publicly available data, including soil quality, geographic location, meteorological data, and rainfall estimates from 46,400 farm households. The methods employed include a combination of advanced models such as Random Forest, Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks.

Focusing on Brazil, a study by Barbosa and Nelson [2016] examines the application of a Support Vector Machine (SVM) classification model. This model is trained on data from farming households in Northeast Brazil to identify household-level characteristics most closely associated with food security and insecurity. The dataset used in this study includes information from approximately 14,000 households, providing a robust sample for analyzing the factors contributing to food insecurity in the region.

While our study aligns with previous research that employs machine learning algorithms to develop predictive

models for food insecurity, it stands out in several key ways. Unlike the studies by Lentz *et al.* [2019] and Christensen *et al.* [2021], our research focuses on predicting food insecurity at the household level, offering a more detailed understanding of the issue. In contrast to Gao *et al.* [2020], we explore a broader range of feature selection methods and classification models to improve prediction accuracy. Furthermore, our approach categorizes food insecurity into multiple levels—ranging from none to light, mild, and severe—providing a more nuanced classification compared to the binary food security/insecurity distinction used by Barbosa and Nelson [2016].

Furthermore, our research provides valuable insights specific to a Global South country, thereby expanding the knowledge base and informing policy decisions in contexts with similar vulnerability profiles. With data from over 33,000 households, our dataset exceeds the size of several datasets used in the referenced studies, offering a strong foundation for our analysis. Through these distinctions, our work not only advances the field of food insecurity research but also highlights the potential of machine learning techniques to address complex social challenges in diverse settings.

5 Data and Methods

In this section, we present an overview of the data sources utilized in this study, along with the transformations applied to prepare the data for analysis. Additionally, we detail the methodology, highlighting the techniques and processes integrated into our proposed solution.

Data Source

The dataset central to this research reflects the conditions of families experiencing extreme social and economic vulnerabilities in the state of Ceará, identified as CMIC families. The CMIC initiative is a permanent public social policy aimed at eradicating extreme child poverty and social vulnerability across the state. It focuses on helping families combat hunger and ensuring food and nutritional security. Through this targeted approach, the CMIC program plays a vital role in supporting vulnerable populations, providing a crucial resource for understanding and addressing the challenges these families face.

Ceará, a northeastern state in Brazil, is characterized by its low economic status, with an average per capita income of just US\$150.00 per month, significantly lower than the national average of US\$643.00. As the 8th most populous state in Brazil, its capital, Fortaleza, ranks as the fourth-largest city in the country, according to the 2022 Brazilian census. Of Ceará's 7 million residents, approximately 25% live in impoverished rural areas, where subsistence agriculture is the primary economic activity. In contrast, around 30% of families participating in the CMIC program reside in rural areas. Although the focus of this study is on Ceará, its implications are broadly significant, particularly as the state is located in a region facing some of Brazil's most severe food insecurity challenges. Government data show that in the Northeast, fewer than half of families (49.7%) have reliable and

full access to food. This statistic underscores the urgency of addressing food insecurity in Ceará and highlights the relevance of this research for understanding and potentially alleviating food insecurity in similar regions.

This study uses data from 33,595 CMIC families, collected between 2020 and 2022 by state-employed researchers who were specifically trained for this task. A standardized form was used to minimize potential input errors [Santana *et al.*, 2023]. By August 2022, approximately half of the targeted families had been surveyed, with the goal of reaching over 120,000 families by the end of the fieldwork in 2023. To qualify as a CMIC family, there must be at least one child aged six years or younger living in the household. The living conditions of these families are marked by housing constructed from makeshift materials, such as reclaimed wood, and the absence of basic amenities, including a bathroom, sanitation facilities, or running water. Additionally, these households have a monthly per capita income of less than US\$16.50 [Santana *et al.*, 2023], highlighting the severe economic and social vulnerabilities they face.

The 29 questions analyzed in this research are grouped into five main categories: Food Insecurity, Food Access, Employment and Income, Housing, and Public Services (as presented in Table 1). An additional category, Location, indicates the city where each surveyed family resides, facilitating the analysis of the dataset's spatial distribution. Although Location data was not used for prediction in this study, it presents a valuable opportunity for future research. Incorporating spatial features could enhance the classification model by providing insights into how food insecurity varies across different geographic areas and potentially uncovering region-specific factors that contribute to the issue.

Table 2 presents the survey questions used as features in the experiments conducted for this study. Due to the proprietary nature of the survey, which is owned by the government of Ceará, the exact wording of these questions cannot be disclosed.

The Food Insecurity component of our study uses a condensed version of the EBIA scale, which is validated in Brazil for assessing food security and endorsed by the Brazilian Ministry of Social Development and Fight Against Hunger [dos Santos *et al.*, 2014]. This scale consists of five questions, with responses ranging from never experiencing the specified aspect of insecurity to experiencing it daily ([Santana *et al.*, 2023], p. 3). Based on the survey responses from 33,595 families, we classified their experiences into four distinct levels of food insecurity, following the EBIA guidelines. Specifically, 10,990 families (33%) were classified as experiencing severe food insecurity, 17,707 families (52%) as experiencing mild food insecurity, 2,165 families (7%) as facing light food insecurity, and 2,733 families (8%) as not experiencing food insecurity.

According to the EBIA scale, families reporting insecurity responses to all five questions were classified as severe; those with three to four responses indicating insecurity were classified as mild; one to two responses were considered light; and those with no signs of insecurity were classified as not experiencing food insecurity. We categorized the light food insecurity and no food insecurity groups as minor labels (representing approximately 15% of the dataset),

Category	Questions	Data type	Themes Investigated
Food Insecurity	1st-5th	Categorical	Whether the family has worried about not having enough food, whether they ran out of food and money to buy some more and whether they skipped meals in the span of the last three months.
Food access	6th-8th	Categorical	Whether Covid-19 impacted their ability to access food and if the family raises animals or plant vegetables in their home.
Employment and Income	9th-14th	Categorical and Numerical	Whether the family receives financial aid, number of people employed, family income.
Public Services	15th-16th	Categorical	Whether the family has used public services offered in community centers to build employable skills, Whether the house is located near recreation and sports infrastructure, or in areas of risk due to armed conflict
Housing	17th-28th	Categorical	Housing tenure, building materials, availability of water, sanitation, trash collection services and public lighting
Location	29	Categorical	City where family is located

Table 1. Feature categories, number of database questions in each category and description.

while the other two were considered major classes (accounting for about 85% of the dataset). Throughout this study, the terms “label” and “class” are used interchangeably to refer to the outcomes we aim to predict.

In addition to the Food Insecurity category, the remaining categories cover a range of topics. The Food Access questions explore alternative strategies that families may use to secure food, such as raising livestock or growing their own crops. Employment-related questions examine whether families have received government vouchers or welfare assistance, the employment status within the household, and the total family income. It is important to note that both the Food Insecurity and Food Access questions played a crucial role in generating the labels for our analysis.

The Housing category examines the infrastructure of the living environment, focusing on the quality and accessibility of essential services such as sanitation and electricity. The Public Services category investigates the social support and developmental opportunities available to family members, including participation in skill-building workshops and access to recreational facilities.

Feature ID	Category	Description
1	Food Insecurity	Whether the family feared that food would run out within the last 3 months
2	Food Insecurity	Whether food ran out before the family could buy some more within the last 3 months
3	Food Insecurity	Whether the family had to eat less due to not having enough money within the last 3 months
4	Food Insecurity	Whether the family ran out of money to maintain a healthy and varied nutrition within the last 3 months
5	Food Insecurity	Whether an adult had to lower food intake or skip meals due to not having enough money to buy food within the last 3 months
6	Food Access	Whether COVID-19 impacted food availability
7	Food Access	Whether the family raises animals for self-consumption
8	Food Access	Whether the family grows food for self-consumption
9	Employment and Income	Whether the family receives government benefits (welfare)
10	Employment and Income	Whether a family member is employed
11	Employment and Income	Number of family members employed
12	Employment and Income	Monthly family income (including government benefits)
13	Employment and Income	Whether someone in the family has previously pursued additional educational training
14	Employment and Income	Whether someone in the family would like to pursue additional educational training
15	Public Services	Availability of parks and recreational facilities near the home
16	Public Services	Availability of cultural activities for children/youth near the home
17	Public Services	Availability of exercise facilities and activities for children/youth near the home
18	Public Services	Whether a family member has participated in activities in the local community center
19	Public Services	Whether a family member has requested social services from the local center
20	Housing	Type of housing ownership/tenure
21	Housing	External wall material of the home
22	Housing	Water source
23	Housing	Whether house has running water
24	Housing	Type of water treatment applied to drinking water
25	Housing	Presence of bathroom in the home
26	Housing	How often does the house have trash pick up
27	Housing	Type of lighting source in the home
28	Housing	Whether the home is located in a violent area
29	Location	City where family lives

Table 2. All feature IDs, Category and description.

Data Source Transformation

As shown in Table 1, the features include both categorical and numerical types, requiring a data transformation step to standardize the feature values within a range from zero to one. The family survey primarily produced **Yes/No** categorical responses, where we assigned a value of **1** for **Yes** and **0** for **No**. Responses marked as **Unknown/Not Informed/Not Given** were coded as **No**.

The primary goal of our coding strategy was to ensure that higher scores accurately reflect a family's vulnerability. However, the initial design of most questions did not directly align with this objective, as most 'Yes' answers were automatically coded as 1 and 'No' answers as 0. As a result, we had to invert the values for some questions to maintain a consistent correlation between the scores and family vulnerability. Additionally, certain questions deviated from the simple Yes/No format, addressing topics such as housing tenure status, energy sources, etc. For these questions, a score of 1 indicates greater vulnerability, while a score of 0 indicates lesser vulnerability.

For example, the question **"In the last 3 months, did you run out of food before having money to buy more?"** offered four possible responses: **"Yes," "No," "Yes, some days,"** and **"Yes, almost every day."** In this case, the last two options, indicating any frequency of running out of food, were assigned a score of **"1"** to denote vulnerability. Consequently, if a categorical response indicates family vulnerability, it is treated as a Yes/No question and assigned a score of **"1"** or **"0"** accordingly. An example of inverted coding is seen in the question **"Is there a bathroom or toilet at home?"** with the responses **"Yes"** and **"No."** Here, the absence of a toilet, which indicates higher vulnerability, was scored as **"1."**

For numerical values, such as **"Number of family members employed?"**, we normalized these values to a range from 0 to 1, ensuring that all feature values were on the same scale.

Methods

Figure 1 outlines the steps of our methodology. Data was initially collected through surveys covering various dimensions of household conditions, including Food Insecurity, Food Access, Employment and Income, Public Services, and Housing. Based on responses related to Food Access and Food Insecurity, each family was assigned to an EBIA category, which served as the primary classification target.

To preprocess the dataset for analysis, the responses were normalized to a range between 0 and 1, ensuring consistency and comparability across all variables. The feature selection process was subsequently performed to identify and eliminate statistically irrelevant features, streamlining the dataset while preserving its essential information. However, the feature selection steps we used in this paper were specifically tailored to our dataset to improve the model performance. For other datasets, these steps might not be necessary — it really depends on the data and the goals of the analysis.

Initially, we tested two feature selection methods: the Chi-square test combined with p-value analysis, and feature

variance evaluation. In this analysis, features with p-values above the significance threshold (set at 0.05 for this study) were considered insufficiently associated with the classification label, suggesting their potential elimination from the dataset. Additionally, we calculated the variance for each feature to identify those exhibiting near-constant variance across most samples, leading to their removal due to lack of variability.

Features directly involved in calculating the label according to the EBIA scale (from the 1st to the 8th feature) and the city (the 29th feature) were excluded from consideration. To handle the imbalance in the dataset, we apply a data balancing technique to ensure fair representation across classes. However, this step can be skipped if class imbalance is not a concern.

Finally, we trained various machine learning classification models, both with and without feature selection, to effectively reduce the number of features required for classification. We employed several feature selection methods in this study in conjunction with the training of classification models, including Select From Model (SFM), Recursive Feature Elimination (RFE), Sequential Feature Selector (SFS), Boruta, and Select K Best (based on Chi-squared tests). Each method was used to evaluate and prioritize features based on their relevance to the classification task. If the optimal model achieves the best performance using all attributes, none of the attributes will be discarded. To optimize the training process and identify the best-performing classification models, we leveraged an AutoML framework. Such automated tool efficiently explores various machine learning algorithms, hyperparameter configurations, and evaluation metrics, streamlining the model selection process.

Finally, we evaluated the classification models to identify the most effective one and conducted an analysis of feature importance. By focusing on the most significant features, we can design new surveys to detect Food Insecurity Levels with fewer features. In addition, this optimized classification model can be used more efficiently to identify food insecurity levels.

Experimental Setup

After refining the feature set through cleaning, normalizing and encoding responses, performing feature selection, and removing statistically irrelevant features, we then split the dataset into training-validation and test subsets, with a split 90% to 10%, respectively. We also implemented k-fold cross-validation within the training-validation subset, further dividing it into 90% for training and 10% for validation. With this setup in place, we proceeded to train various classification models, using the distinct feature selection methods outlined above.

In the experiments, we use two techniques from the Imbalanced Learn library to balance the dataset, SMOTE [Chawla et al., 2002] and SMOTE-Tomek [Tomek, 1976]. We also use H2O AutoML [LeDell and Poirier, 2020], an open source, machine learning software commonly used for automating the machine learning training. The current version of H2O AutoML trains, beyond others, the following models: XGBoost GBM (Gradient Boosting Machine), a fixed grid of

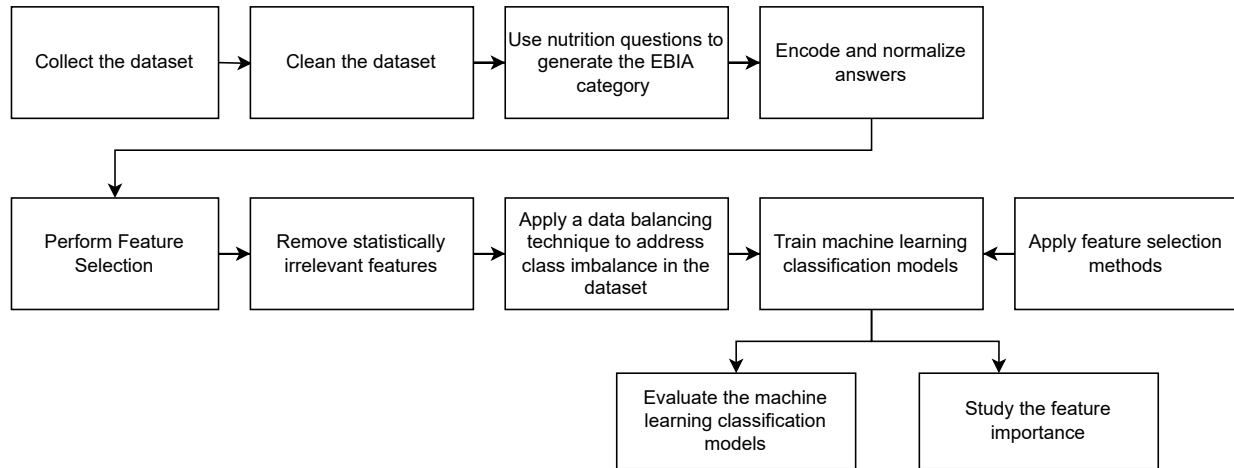


Figure 1. Methodology developed in study.

Generalized Linear Model with regularization (GLM), Random Forest (RF) and its variants, Stacked Ensemble and a random grid of deep neural networks. Table 3 presents the H2O parameters settings for executing the process to test all models. All other parameters that were not included run with default values are described in the H2O documentation.

Parameter	Value
max_runtime_secs	0
seed	42
nfolds	10

Table 3. H2O parameters used

Those tree-based models are used to select the features with the strongest relationship to the dependent or target variable. Table 4 presents the parameters and estimators used in each method of feature selection. As explained before, to validate and estimate the performance of the models, we used K-fold cross-validation and for the test was used the train_test_split function from the scikit-learn with 10% of the dataset dedicated to the test. The Grid Search method used was Cartesian Grid Search, that is the default grid search method in H2O, it exhaustively searches over all possible combinations of the model's hyperparameters.

We trained a variety of algorithms, resulting in a diverse set of potential models. This approach enabled us to explore different combinations, helping refine and develop a more accurate final model to effectively address the research questions. All experiments were executed with Python 3.8 on Google Colab, using a Linux platform with, an Intel(R) Xeon(R) CPU, 2.20GHz, 12GB RAM, Nvidia K80/T4.

6 Results

This section outlines the results obtained from our experiments and how they align with the research questions.

6.1 Feature Engineering

The first step in our analysis involved feature selection strategies to identify and exclude features unrelated to the target

Method	Estimators	Parameter	Value
Select From Model	Logistic Regression	max_iter	100000
	Logistic Regression CV		
	Passive Aggressive Classifier		
	Perceptron		
	Ridge Classifier		
	Ridge Classifier CV		
	SGD Classifier		
	SGD OneClass		
	SVM		
Boruta	Random Forest Regressor	n_jobs	-1
		max_depth	5
		n_estimators	auto
Sequential Feature Selection	Ridge Classifier (Forward)	random_state	0
		cv	10
	Ridge Classifier (Backward)	max_iter	100000
Select K Best	Logistic Regression (Forward)		
	Logistic Regression (Backward)		
Recursive Feature Elimination	chi2	k	[1, 17]
	f_classif		
	Decision Tree Classifier	n_feature_to_select	[1, 17]

Table 4. Feature selection methods and respective estimators and parameters

label of food insecurity. Feature selection can be seen as a preprocessing step for an estimator which helps reduce the dimensionality of a dataset and allows for the identification of the most relevant features through univariate statistical tests.

6.1.1 Chi-Square and Significance Tests

First we used Chi-square (χ^2) tests and p-value analysis. Features with $\rho \geq 0.05$ were deemed irrelevant and excluded from the training set [Howell, 2011]. This step was important to explore the relationship between features F and the label Y in our dataset D_F and address Research Question 1 about which features can predict food insecurity. Here, Y represents the four levels of food insecurity: “Severe”, “Moderate”, “Mild”, and “Without Food Insecurity”. Table 5 presents the results of this feature selection process, highlighting excluded features in bold.

Feature ID	χ^2	ρ - value
1	13151.061500	0.000000000
2	6486.323340	0.000000000
3	6031.457090	0.000000000
4	13268.402300	0.000000000
5	3971.765830	0.000000000
6	2300.500410	0.000000000
7	23.105797	0.000038384
8	43.737673	0.000000002
9	4.606051	0.203023593
10	125.972363	0.000000000
11	23.892548	0.000026304
12	20.243117	0.000151145
13	6.014177	0.110922495
14	36.776572	0.000000051
15	56.527859	0.000000000
16	12.654843	0.005445719
17	58.083084	0.000000000
18	30.151232	0.000001283
19	15.589467	0.001376307
20	25.064534	0.000014968
21	167.276171	0.000000000
22	32.932614	0.000000333
23	141.781679	0.000000000
24	103.904725	0.000000000
25	45.427670	0.000000001
26	26.979846	0.000005945
27	1.037621	0.792149893
28	10.052388	0.018126029
29	137.733032	0.000000000

Table 5. χ^2 and ρ -values computed for all features. ρ -values ≥ 0.05 highlighted in bold.

We identified that the 9th, 13th, and 27th features have ρ -values exceeding the significance threshold, indicating a lack of association with the label and suggesting their exclusion from the dataset. Furthermore, after evaluating the variance of each feature, we observed values below 0.07, indicating that approximately 93% of all samples have the same value. This analysis confirms that these features are quasi-constant and supports their removal. The eliminated features correspond to questions regarding the family’s receipt of welfare benefits, their level of educational attainment, and the type of lighting source used in their homes. Given the vulnerable status of CMIC families, it is plausible that many share similar circumstances related to welfare receipt and limited educational opportunities, among other factors.

Observations from Table 5 show that the χ^2 values for the

1st through 6th features are significantly higher than those for the other features. This outcome is expected, as the food insecurity level, which serves as our label, is derived from the values of the first eight features, including the 1st through 6th. These features were then excluded from the dataset D_F . The 29th characteristic, which indicates the city where the family resides, was also excluded from our analysis, as not all cities in the Ceará state were surveyed.

After applying this test, we retained 17 features: 11 from the “Housing” category, 4 from “Employment and Income,” and 2 from “Public Services.” These features formed the basis for subsequent model training.

6.2 AutoML model results for different sampling techniques

Before training the model, we performed a stratified split of the dataset D_F into training-validation (90%) and test (10%) sets. This stratification ensures that the label distribution in the sample reflects that of the original dataset. Specifically, our data set comprises 33,595 responses on four labels: 33% classified as severe food insecurity, 7% as light food insecurity, approximately 52% as mild food insecurity, and 8% as no food insecurity. The stratification ensures that the labels for severe, mild, light and no food insecurity are represented by approximately 33%, 52%, 7%, and 8%, respectively, in the training and test sets. Following this, we performed an additional stratified split of the training-validation set into 90% for training and 10% for validation.

Addressing the inherent class imbalance of the dataset was crucial to improving model performance. To this end, we tested two combined sampling techniques: SMOTE and SMOTE-Tomek [Batista *et al.*, 2003]. Although we also tested using only SMOTE for oversampling, this approach did not yield better results.

6.3 Multi-Label Classification Results

To evaluate the impact of removing features based on a ρ -value ≥ 0.05 , as determined by the χ^2 approach, we used AutoML to select the best model for the imbalanced dataset. We excluded only the features used to compute the food insecurity level — namely, the first eight features and the 29th feature (related to the city) from Table 5 — leaving us with 20 features. Using the same setup, we stratified the dataset D_F into training-validation (90%) and test (10%) sets. We also performed an additional stratified split, dividing the training-validation set into 90% for training and 10% for validation. As shown in Figure 2, XGBoost emerged as the best model. However, it failed to outperform both our imbalanced model with fewer features (17) and the balanced model, which also had fewer features (17). Consequently, we conclude that features with a p-value greater than or equal to 0.05 are negligible and can be excluded during the model training process.

We relied on AutoML to identify the optimal model for our classification task, utilizing the balanced dataset achieved through the SMOTE-Tomek technique. With the refined set of features, we used AutoML to identify the best model from D_F for predicting Y . The results are presented in Figures 2 when considering 4 classes and Figure 3 when considering 2

classes. We will continue the analysis based on the results of the models trained for 4 classes and revisit the performance of the models for 2 classes in Section 6.4.

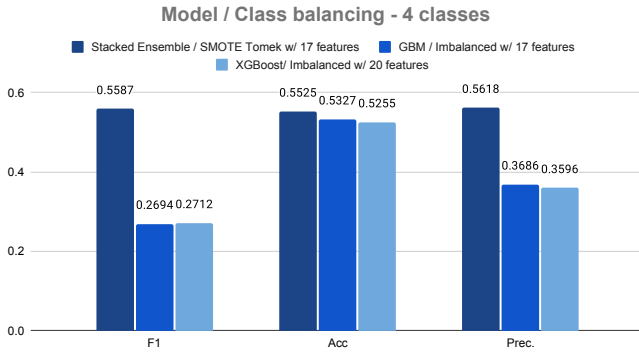


Figure 2. Results for the best machine learning model selected by the AutoML when considering four classes.

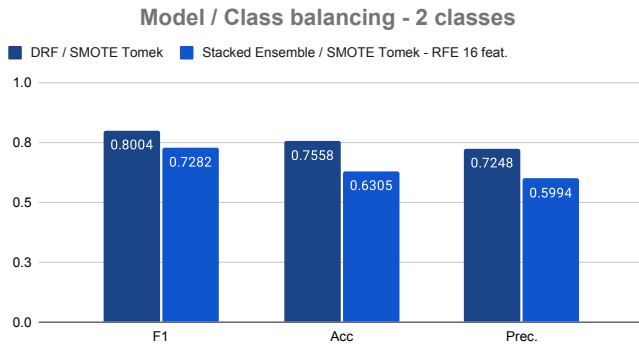


Figure 3. Results for the best machine learning model selected by the AutoML when considering two classes.

For the imbalanced dataset, the best-performing model identified by AutoML was the Gradient Boosting Machine (GBM), which achieved an accuracy of approximately 0.53. However, due to the GBM model's difficulty in accurately predicting the smaller classes (severe, light, and no food insecurity), its F1-score was relatively low, around 0.26.

When using SMOTE-Tomek to balance the dataset, AutoML selected a Stacked Ensemble as the best-performing model. This model achieved an accuracy of 0.55 and an F1-score of 0.55.

	Pred. Severe	Pred. Light	Pred. Mild	Pred. W/o FI
True Severe	5250	2107	3683	1947
True Light	897	9186	1316	2222
True Mild	2537	1726	6973	1751
True W/o FI	945	2379	1649	8647

Table 6. Confusion matrix of the best model found for the 4-class scenario, i.e., Stacked Ensemble handling the imbalanced dataset via SMOTE-Tomek. Predicted as Pred. and Without as W/o

Table 6 presents the confusion matrix of our best model for the 4-class scenario, utilizing a Stacked Ensemble approach. The imbalanced dataset was handled using SMOTE-Tomek, and feature engineering techniques were applied. Figure 4 shows the Recall and Precision values for each class. After the over- and under-sampling provided by SMOTE-Tomek, the labels exhibit almost the same number of instances. The recall (fraction of relevant instances retrieved) for Mild and

Severe food insecurity was approximately 0.53 and 0.40, respectively. Meanwhile, the recall for Light and No food insecurity was around 0.67 and 0.63, respectively.

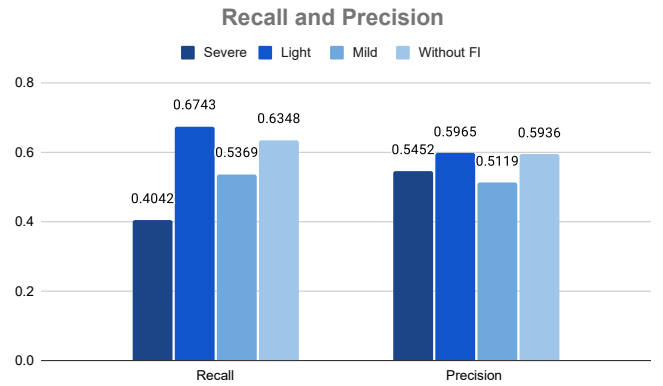


Figure 4. Recall and Precision of the best model found on RQ1, i.e., the Stacked Ensemble handling the imbalanced dataset via SMOTE-Tomek

6.3.1 Other feature selection strategies and their impact on model prediction

We also evaluated various feature selection methods available in Scikit-learn [Pedregosa *et al.*, 2011] to find whether it was possible to reduce features but maintain or improve model accuracy. This was part of addressing Research Question 2.

Several feature selection techniques were employed in this stage to find a model with reduced features: (1) Select K Best, which combines univariate statistical tests (such as Chi-square and Pearson correlation) with the selection of the top K features from the dataset X based on the statistical relationship between X and y ; (2) Select from Model (SFM), which uses a machine learning model to estimate the weight of each feature's importance, selecting features based on a predefined weight threshold; (3) Recursive Feature Elimination (RFE), which also employs a machine learning model to identify the most important features by recursively eliminating the least important ones after each training phase, until the desired number of features is reached; (4) Sequential Feature Selection (SFS), which adds (forward selection) or removes (backward selection) features to form feature subsets, with the estimator choosing the best feature to add or remove at each stage based on cross-validation scores; and (5) Gradient Boosting, a classification model that leverages an ensemble of weaker classifiers.

In terms of results, the best AutoML model was a Stacked Ensemble that used the RFE feature selection algorithm. However, this model did not reduce the number of features, opting to retain all seventeen features identified in the experiments for (RQ1).

In short, the initial feature selection method using the combined Chi-square (χ^2) test and p -values, was the most suitable method to create a parsimonious model with good accuracy.

6.4 Binary Classification Model Results

To simplify the problem, we consolidated the four food insecurity classes into two: severe and non-severe food insecurity. This binary classification approach was designed to evaluate whether a simpler label structure could improve model performance. The binary labels were defined by combining the "mild," "light," and "without FI" categories into a single "non-severe FI" class, while retaining "severe FI" as a separate class. Approximately 33% of the dataset was labeled as severe FI, with the remainder classified as non-severe FI.

Returning to the analysis of Figure 3, we have already highlighted the importance of balancing the dataset. Next, we aim to investigate the relevance of feature selection and its impact on model performance in a simpler problem, specifically the 2-class scenario. AutoML identified the Stacked Ensemble as the optimal model for the 2-class scenario when feature selection was combined with the training of classification models. The Stacked Ensemble achieved an accuracy of 0.63 and an F1-score of 0.72. Notably, this model excluded the feature related to participation in community center activities, retaining 16 features in total.

However, the best-performing model retained all seventeen features. In summary, the initial feature selection method, using the combined Chi-square (χ^2) and ρ -values, proved to be the most effective method for developing a concise model with strong accuracy. The best model generated by AutoML was a Distributed Random Forest (DRF), which achieved an accuracy of approximately 0.75 and an F1-score of 0.80. This result outperforms all other models we tested. The confusion matrix for this model is shown in Table 7.

	Predicted Severe	Predicted Non-Severe
True Severe	5527	3351
True Non-Severe	1007	8612

Table 7. Confusion matrix of the best model found for the 2-class scenario with Distributed Random Forest handling the imbalanced dataset via SMOTE-Tomek.

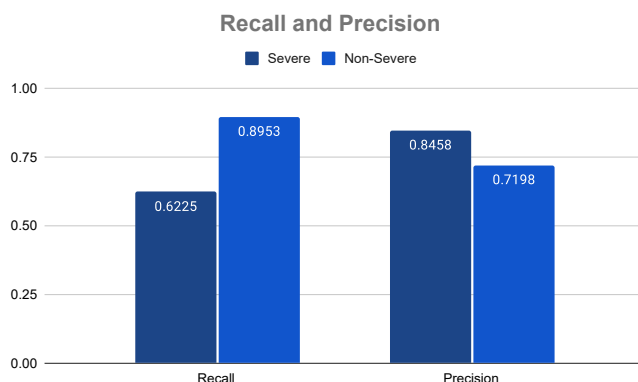


Figure 5. Recall and Precision of the best model found with Distributed Random Forest handling the imbalanced dataset via SMOTE-Tomek

After the over- and under-sampling provided by SMOTE-Tomek, the labels exhibit nearly the same number of instances. The recall for severe and non-severe food insecurity are approximately 0.62 and 0.89, respectively (Figure 5).

Feature Description	Category	Percentage Importance
Monthly family income (including government benefits)	Employment and Income	41.45
Type of housing ownership/tenure	Housing	5.31
Whether someone in the family would like to pursue additional educational training	Employment and Income	5.14
Water source	Housing	5.06
Whether a family member has participated in activities in the local community center	Public Services	4.68
How often does the house have trash pick up	Housing	4.43
Availability of exercise facilities and activities for children/youth near the home	Public Services	4.43
Availability of parks and recreational facilities near the home	Public Services	4.35
External wall material of the home	Housing	4.22
Whether house has running water	Housing	3.94
Type of water treatment applied to drinking water	Housing	3.94
Presence of bathroom in the home	Housing	3.07
Number of family members are employed	Employment and Income	2.69
Whether the home is located in a violent area	Housing	2.65
Whether a family member is employed	Employment and Income	2.12
Whether there are cultural activities for children/youth near the home	Public Services	1.82
Whether a family member has requested social services from the local center	Public Services	0.70

Table 8. Feature Rank by best model for balanced 2 classes without feature selection (Distributed Random Forest).

6.5 Feature Importance Analysis

Feature importance was assessed for both multi-label and binary classification tasks. For the multi-label task, SHAP values highlighted "Employment and Income" and "Housing" features as the most influential. For the binary task, H2O's improvement metric provided similar results. Figure 6 illustrates the mean absolute SHAP values for the most important features in the balanced Stacked Ensemble model. Regarding the ranking of the most important features, "Employment and Income" and "Housing" emerged as the two most significant categories for determining the level of food insecurity (Table 8). The "Employment and Income" features contribute 51.40% to the total feature importance, while "Housing" accounts for 32.62%, and "Public Services" makes up 15.98%.

H2O computes feature importance by analyzing the squared error before and after a split using a particular variable during the training of the DRF model. This difference is termed "improvement". H2O then calculates the total feature importance by summing the improvement values for each feature that contributed to a split. The total importance for all

features is then scaled between 0 and 1. These scaled values are further converted into percentage importance, ensuring the total sum equals 100%. The percentage importance for each feature is presented in Table 8.

Among these features, family income is nearly eight times more important than the second most significant feature, a finding that aligns with other studies exploring the relationship between these variables [Gao *et al.*, 2020]. Research by Nord [2007] indicates that a large portion of the US population experiencing food insecurity is also living below the poverty line, which is further correlated with high unemployment rates. Similarly, in Brazil, Reis [2012] demonstrated that a lack of financial resources for food is associated with poorer nutritional outcomes and health indicators. Other studies about housing tenure point to the importance of adequate housing and ownership to family nutrition [Lee *et al.*, 2021]. Specifically, Fletcher *et al.* [2009] found that a \$500 increase in annual rental costs correlates with a nearly 3% increase in food insecurity rates. Housing instability, which includes missed rent or mortgage payments, overcrowding, frequent moves, evictions, or homelessness, is strongly linked to food insecurity, as stable housing improves access to services and facilitates the building of supportive social networks [King, 2018]. Our findings are consistent with existing literature and contribute to the growing body of research on food insecurity, particularly within the Brazilian context.

Features from the "Employment and Income" and "Housing" categories are the most influential for the classification model. This analysis was conducted using only 10% of the dataset due to limitations in H2O, which prevented us from processing the entire training set. We will continue to explore ways to overcome these limitations in SHAP value calculations when working with larger datasets like ours.

It is important to emphasize the uniqueness of our dataset, which was derived from the state of Ceará. This dataset is the only one in Brazil that combines detailed nutrition data with a broad range of social variables, providing a comprehensive view of the factors influencing food insecurity. While Ceará has its distinct characteristics, it shares many socio-economic vulnerabilities with other Brazilian states, especially the 16 states in the North and Northeast regions. Notably, 53% of residents in the Northeast live below the poverty line, and 13.3% in the North earn less than USD 1.2 per day. These similarities suggest that, although the dataset is specific to Ceará, our findings may have relevance and applicability to broader contexts across Brazil, particularly in regions facing similar challenges.

The results show that predicting the four levels of food insecurity is complex. This complexity is evident when we compare it with the simpler model that predicts whether insecurity is severe or not. We find our results to be reasonable, as, to the best of the authors' knowledge, no existing software currently computes food insecurity levels using features unrelated to nutrition. Furthermore, even the Without FI and Light FI labels, despite having fewer samples in the original dataset, achieve competitive recall when compared to the major classes.

7 Final Remarks

Food insecurity represents a major global challenge, particularly in developing countries. In Brazil, the focus of this study, three out of every ten people face uncertainty about their ability to secure their next meal [Organization *et al.*, 2021]. While the pandemic has significantly worsened global hunger, social inequality and the lack of effective public policies also play critical roles in exacerbating the problem. As Foini *et al.* [2022] emphasize, obtaining accurate and timely data to understand the scope of this challenge is crucial. They argue, "An essential step towards achieving hunger reduction is to have access to frequent, up-to-date information on the status of food insecurity in countries facing humanitarian crises, and to estimates of where and when the situation is likely to improve or deteriorate, in order to allow for informed and timely decision-making on resource allocation and on relevant policies and programmes."

However, measuring food insecurity and generating accurate statistics to inform policymakers presents significant logistical and financial challenges [Nica-Avram *et al.*, 2020], and Brazil is no exception. The 2020 Federal Census was delayed until 2022 due to the pandemic and faced a budget cut exceeding 90%, just as surveys were ready for distribution. When the census finally began, it was revealed that these budgetary constraints led to the elimination of 25% of the planned questions [Collado *et al.*, 2024]. This situation highlights the need for alternative, data-driven, and predictive strategies to address the operational challenges of surveying populations, particularly in Global South countries, which often cover vast territories. This study addresses these challenges by using an automated approach to analyze food insecurity through more universally applicable household-level questions.

The methodology consisted of several key steps, utilizing AutoML techniques to explore three research questions. First, we implemented feature selection strategies to address RQ1, which aimed to determine whether it was feasible to develop a model for predicting food insecurity. The resulting model achieved an accuracy of 0.55, demonstrating that meaningful insights could be derived from a reduced set of features. Out of the original 29 features, 17 were retained, relating to public services (5 features), employment and income (4 features), and housing (8 features). For RQ2, which explored the potential for future simplification through AutoML, it was observed that the best-performing AutoML model chose to retain all 17 features, indicating no reduction in feature count. Finally, RQ3 examined whether consolidating the classification into two categories—severe and non-severe food insecurity—would improve the model's performance. The optimal model, identified by AutoML as a Distributed Random Forest applied to a balanced dataset, utilized 16 features and achieved an accuracy of approximately 0.75, with an F1-score of 0.80. Notably, the "Employment and Income" and "Housing" categories were found to be the most influential in determining the level of food insecurity.

In summary, automated solutions, particularly machine learning (ML) models, have proven effective in forecasting and predicting food insecurity. These models can play a crucial role in addressing the challenge of measuring food inse-

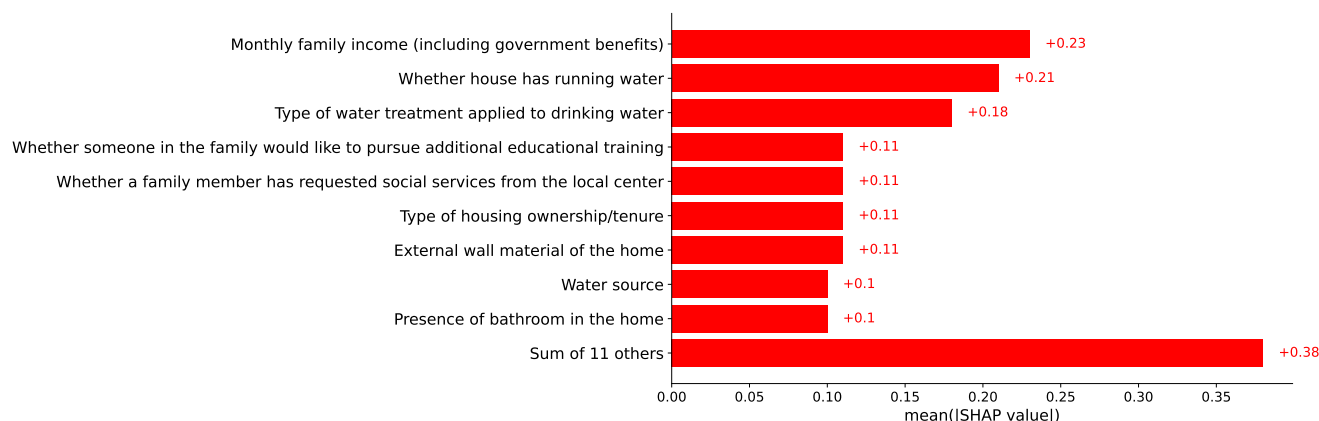


Figure 6. Feature importance study of the best model found with 10% of balanced dataset due to H2O limitation.

curity, offering valuable insights that can inform policymaking, especially in the absence of government-led measurements.

It is important to recognize that having adequate data alone does not solve all challenges. The feature importance analysis identifies monthly family income, including government benefits, as a critical factor, highlighting the central role of financial stability in this study's context. This finding underscores the significant social vulnerability of the families involved and reinforces the need for effective policies to reduce hunger, which must include economic interventions that address its root causes. Likewise, the importance of features related to housing and community services, such as participation in local community center activities, access to exercise facilities, parks, recreational spaces, and cultural events, emphasizes the value of a supportive, resource-rich community environment. This suggests a need for increased investment in public infrastructure and community resources to address broader social determinants, aiming to prevent food insecurity in the long term.

Classifying food insecurity as severe or non-severe yielded satisfactory results, though there remains room for improvement. In contrast, predicting the four levels of food insecurity based on socioeconomic factors is more complex and demands further investigation and research. For future work, we propose exploring the use of neural networks with more robust frameworks that provide greater flexibility in designing network architectures. Additionally, employing explanation techniques to understand how models distinguish among the four levels of food insecurity could offer valuable insights to better address this challenge.

Declarations

Acknowledgements

We acknowledge the work of social workers from the Government of Ceará for their thorough efforts in collecting data on CMIC families.

Funding

The research reported in this work was supported by the Ceara Foun-

ation for Research Support (FUNCAP), under project "Big Data Platform to Accelerate the Digital Transformation of Ceará State" [grant number 04772551/2020] and "Citizen Platform" [grant number 04772314/2020].

Authors' Contributions

Ticiania L. Coelho da Silva: Methodology, Conceptualization, Supervision, Project administration, Funding acquisition, Writing - Review and Editing. Lara Sucupira Furtado: Conceptualization, Supervision, Writing - Review and Editing. Guilherme Sales Fernandes: Software, Methodology, Writing - Review and Editing. José A. Fernandes de Macêdo: Funding acquisition, Writing - Review and Editing. Lívia Almada Cruz: Methodology, Conceptualization, Writing - Review and Editing. Regis Pires Magalhães: Methodology, Conceptualization, Writing - Review and Editing. Laécia Gretha Amorim Gomes: Data Curation.

Competing interests

The authors declare no conflict of interest. The founders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Availability of data and materials

Due to confidentiality agreements, supporting data can only be made available to bona fide researchers subject to a non-disclosure agreement

References

- Barbosa, R. M. and Nelson, D. R. (2016). The use of support vector machine to analyze food security in a region of Brazil. *Applied Artificial Intelligence*, 30(4):318–330. DOI: 10.1080/08839514.2016.1169048.
- Batista, G. E., Bazzan, A. L., Monard, M. C., et al. (2003). Balancing training data for automated annotation of keywords: a case study. *Wob*, 3:10–18. Available at: <https://www.inf.ufrgs.br/maslab/masbio/papers/balancing-training-data-for.pdf>.

- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140. DOI: 10.1007/BF00058655.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. DOI: 10.1023/A:1010933404324.
- Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168. ACM. DOI: 10.1145/1143844.1143865.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. DOI: 10.48550/arXiv.1106.1813.
- Christensen, C., Wagner, T., and Langhals, B. (2021). Year-independent prediction of food insecurity using classical and neural network machine learning methods. *AI*, 2(2):244–260. DOI: 10.3390/ai2020015.
- Collado, L. F., Leichsenring, A. R., and Moutian, A. G. (2024). A saga do censo demográfico brasileiro de 2020. *Boletim de Políticas Públicas/OIPP No16 agosto/2021*, 29. Available at: https://sites.usp.br/boletimoipp/wp-content/uploads/sites/823/2021/10/Collado_Leichsenring_Moutian_agosto_2021.pdf.
- Costa, N. S., Santos, M. O., Carvalho, C. P. O., Assunção, M. L., and Ferreira, H. S. (2017). Prevalence and factors associated with food insecurity in the context of the economic crisis in brazil. *Current Developments in Nutrition*, 1(10):e000869. DOI: 10.3945/cdn.117.000869.
- Deléglise, H., Bégué, A., Interdonato, R., d’Hôtel, E. M., Roche, M., and Teisseire, M. (2020). Linking heterogeneous data for food security prediction. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 335–344. Springer. DOI: 10.1007/978-3-030-65965-3_2.
- Deléglise, H., Interdonato, R., Bégué, A., d’Hôtel, E. M., Teisseire, M., and Roche, M. (2022). Food security prediction from heterogeneous data combining machine and deep learning methods. *Expert Systems with Applications*, 190:116189. DOI: 10.1016/j.eswa.2021.116189.
- dos Santos, L. P., Lindemann, I. L., dos Santos Motta, J. V., Mintem, G., Bender, E., and Gigante, D. P. (2014). Proposta de versão curta da escala brasileira de insegurança alimentar. *Revista de Saúde Pública*, 48(5):783–789. Available at: <https://www.scielo.br/j/rsp/a/m4WdfKXNhLfXtc3b8fpQg6D/?lang=pt&format=pdf>.
- Felker-Kantor, E. and Wood, C. H. (2012). Female-headed households and food insecurity in brazil. *Food Security*, 4(4):607–617. DOI: 10.1007/s12571-012-0215-y.
- Fernández, A., García, S., del Jesus, M. J., and Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, 159(18):2378–2398. DOI: 10.1016/j.fss.2007.12.023.
- Fletcher, J. M., Andreyeva, T., and Busch, S. H. (2009). Assessing the effect of changes in housing costs on food insecurity. *Journal of Children and Poverty*, 15(2):79–93. DOI: 10.1080/10796120903310541.
- Foini, P., Tizzoni, M., Martini, G., Paolotti, D., and Omodei, E. (2022). On the forecastability of food insecurity. *medRxiv*, pages 2021–07. DOI: 10.1038/s41598-023-29700-y.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232. DOI: 10.1214/aos/1013203451.
- Furtado, L. S. and Furtado, L. S. (2021). Urban collectives and insurgency to fight covid-19: an analysis of social media content. *Oculum Ensaios*, 18:1–21. DOI: 10.24220/2318-0919v18e2021a5136.
- Gao, C., Fei, C. J., McCarl, B. A., and Leatham, D. J. (2020). Identifying vulnerable households using machine learning. *Sustainability*, 12(15). DOI: 10.3390/su12156002.
- Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *International Journal of Neural Systems*, 8(02):209–218. DOI: 10.1142/s0129065797000227.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42. DOI: 10.1145/3236009.
- 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. DOI: 10.1016/j.eswa.2016.12.035.
- Hanif, A., Beheshti, A., Benatallah, B., Zhang, X., Habiba, Foo, E., Shabani, N., and Shahabikargar, M. (2023). A comprehensive survey of explainable artificial intelligence (xai) methods: Exploring transparency and interpretability. In *International Conference on Web Information Systems Engineering*, pages 915–925. Springer. DOI: 10.1007/978-981-99-7254-8_71.
- He, X., Zhao, K., and Chu, X. (2021). Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622. DOI: 10.48550/arXiv.1908.00709.
- Howell, D. C. (2011). Chi-square test: analysis of contingency tables. In *International Encyclopedia of Statistical Science*, pages 250–252. Springer. Available at: <https://www.uvm.edu/~statdhtx/methods8/Supplements/Chi-Square-Folder/ChiSquareTests.pdf>.
- King, C. (2018). Food insecurity and housing instability in vulnerable families. *Review of Economics of the Household*, 16(2):255–273. DOI: 10.1007/s11150-016-9335-z.
- LeDell, E. and Poirier, S. (2020). H2O AutoML: Scalable automatic machine learning. *7th ICML Workshop on Automated Machine Learning (AutoML)*. Available at: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.
- Lee, C. Y., Zhao, X., Reesor-Oyer, L., Cepni, A. B., and Hernandez, D. C. (2021). Bidirectional relationship between food insecurity and housing instability. *Journal of the Academy of Nutrition and Dietetics*, 121(1):84–91. DOI: 10.1016/j.jand.2020.08.081.
- Lentz, E. C., Michelson, H., Baylis, K., and Zhou, Y. (2019). A data-driven approach improves food insecurity crisis prediction. *World Development*, 122:399–409. DOI: 10.1016/j.worlddev.2019.06.008.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified ap-

- proach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.. DOI: 10.48550/arXiv.1705.07874.
- Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition. Available at: <https://christophm.github.io/interpretable-ml-book>.
- Neri, M. (2022). Insegurança alimentar no brasil: pandemia, tendências e comparações internacionais. *Rio de Janeiro: FGV Social*. Available at: <https://hdl.handle.net/10438/32406>.
- Nica-Avram, G., Harvey, J., Goulding, J., Lucas, B., Smith, A., Smith, G., and Perrat, B. (2020). Fims: Identifying, predicting and visualising food insecurity. In *Companion Proceedings of the Web Conference 2020*, pages 190–193. DOI: 10.1145/3366424.3383538.
- Nord, M. (2007). Characteristics of low-income households with very low food security: an analysis of the usda gpra food security indicator. *USDA-ERS Economic Information Bulletin*, (25). Available at: <https://www.ers.usda.gov/publications/pub-details?pubid=44173#download>.
- Organization, W. H. et al. (2021). *The State of Food Security and Nutrition in the World 2021: Transforming food systems for food security, improved nutrition and affordable healthy diets for all*, volume 2021. Food & Agriculture Org. Available at: <https://openknowledge.fao.org/server/api/core/bitstreams/1c38676f-f5f7-47cf-81b3-f4c9794eba8a/content>.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. Available at: <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>.
- Reis, M. (2012). Food insecurity and the relationship between household income and children's health and nutrition in brazil. *Health economics*, 21(4):405–427. DOI: 10.1002/hec.1722.
- Saarela, M. and Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, 3(2):272. DOI: 10.1007/s42452-021-04148-9.
- Sammut, C. and Webb, G. I. (2017). *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated. DOI: 10.1007/978-1-4899-7687-1.
- Santana, O. M. M. L. d., Sousa, L. V. d. A., Lima Rocha, H. A., Correia, L. L., Gomes, L. G. A., Aquino, C. M. d., Rocha, S. G. M. O., Araújo, D. A. B. S., Soares, M. D. d. A., Machado, M. M. T., et al. (2023). Analyzing households' food insecurity during the covid-19 pandemic and the role of public policies to mitigate it: evidence from ceará, brazil. *Global Health Promotion*, 30(1):53–62. DOI: 10.1177/17579759221107035.
- Shapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651–1686. DOI: 10.1214/aos/1024691352.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117. DOI: 10.1016/j.neunet.2014.09.003.
- Segall-Corrêa, A. M. and Marin-Leon, L. (2009). A segurança alimentar no brasil: proposição e usos da escala brasileira de medida da insegurança alimentar (ebia) de 2003 a 2009. *Segurança Alimentar e Nutricional*, 16(2):1–19. DOI: 10.20396/san.v16i2.8634782.
- Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. *Handbook of statistics*, 24:303–329. DOI: 10.1016/S0169-7161(04)24011-1.
- Tang, J., Alelyani, S., and Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37. Available at: https://www.cse.msu.edu/~tangjili/publication/feature_selection_for_classification.pdf.
- Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772. DOI: 10.1109/TSMC.1976.4309452.
- Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press. Book.