Analysis of Socioeconomic and Anthropometric Data via Tree-based Models: Evidence for Policies Against Hunger

João Gabriel Soares Ferreira[1], **César Lincoln Cavalcante Mattos**[1], **Antonio Rafael Braga**[2,3], **Danielo G. Gomes**[3]

[1]Departamento de Computação – Centro de Ciências,
Universidade Federal do Ceará (UFC), Fortaleza – CE

[2]Redes de Computadores – Campus Quixadá,
Universidade Federal do Ceará (UFC), Quixadá – CE

[3]Departamento de Engenharia de Teleinformática – Centro de Tecnologia,
Universidade Federal do Ceará (UFC), Fortaleza – CE

gabrielsoares@alu.ufc.br, cesarlincoln@dc.ufc.br,
rafaelbraga@ufc.br, danielo@ufc.br

Resumo. *A fome e a insegurança alimentar persistem como desafios tanto em escala regional quanto global, sendo agravados pela limitada disponibilidade e acessibilidade a dados sobre essas questões. No entanto, dados socioeconômicos a nível individual (ou de agregado familiar) estão disponíveis, e são coletados, em todo o Brasil. Nesse sentido, este trabalho propõe usar modelos de aprendizado de máquina capazes de predizer indicadores antropométricos ligados à fome, a partir de dados socioeconômicos. Os dados em questão foram extraídos automaticamente da plataforma CECAD (Consulta, Seleção e Extração de Informações do CadÚnico). Os indicadores antropométricos (baixo peso para altura, baixo peso para idade e baixa altura para idade) foram coletados no SISVAN (Sistema de Vigilância Alimentar e Nutricional). Os experimentos focaram-se em modelos baseados em árvore de decisão (Random Forest, Gradient Boosting, XGBoost, LightGBM e CatBoost). Na avaliação dos modelos, todos os municípios brasileiros foram usados para o treinamento, com a exceção daqueles do estado do Ceará, que foram separados para teste. Os melhores modelos obtiveram resultados promissores na tarefa de predição, especialmente para o indicador de baixa altura para a idade, em que o erro percentual chegou a* 22%.

Abstract. *Hunger and food insecurity are regional and global problems we still face, being intensified by the lack of broad access to data on these issues. However, socioeconomic data at the individual (or household) level are available and collected all over Brazil. In that sense, this work proposes to use machine learning models capable of predicting anthropometric indicators related to hunger, based on socioeconomic data. The data in question was automatically extracted from the CECAD platform (Query, Selection and Extraction of Information from CadÚnico). The anthropometric indicators (low weight for height, low*

weight for age and low height for age) were collected from SISVAN (Food and Nutrition Surveillance System). The experiments focused on decision tree-based models (Random Forest, Gradient Boosting, XGBoost, LightGBM and CatBoost). All Brazilian municipalities were used for model training, with the exception of those in the state of Ceará, which were separated for testing. The best models obtained promising results in the prediction task, especially for the low height-for-age indicator, where the percentage error reached 22%.

1. Introdução

Em 2022, 21,1 milhões de brasileiros se encontravam em insegurança alimentar grave, caracterizada por estado de fome [World Health Organization 2023]. A insegurança alimentar é caracterizada pelo acesso inadequado a alimentos seguros e nutritivos, essenciais para uma vida ativa e saudável, bem como para um crescimento e desenvolvimento normais. Tal condição pode decorrer da indisponibilidade de alimentos e/ou da insuficiência de recursos para sua aquisição [Food and Agriculture Organization of the United Nations 2023]. É importante ressaltar que a deficiência de nutrientes acarreta uma série de consequências para a saúde pública, incluindo prejuízos cognitivos, atraso no desenvolvimento físico e a predisposição a doenças [Sousa and Diniz 2024].

Indicadores antropométricos são medidas físicas utilizadas para avaliar características corporais e morfológicas de uma pessoa. Eles incluem medições como altura, peso, circunferência da cintura, dobras cutâneas e índice de massa corporal (IMC). Esses indicadores são frequentemente usados em estudos de saúde pública, nutrição e medicina para avaliar o crescimento, o desenvolvimento e o estado nutricional das pessoas [Aguiar et al. 2023]. Eles fornecem informações importantes sobre a composição corporal, o estado de saúde e o risco de certas condições, como desnutrição, obesidade e outras doenças relacionadas ao peso e à nutrição. Os indicadores socioeconômicos representam outra maneira de avaliar a insegurança alimentar, ou seja, a falta de acesso à renda, bens e serviços, além das mudanças nutricionais, também causam impactos no desenvolvimento psicomotor dos indivíduos [Morais et al. 2014].

Diversas pesquisas disponibilizam somente dados sumarizados a nível nacional, regional ou estadual. Bases de dados que contêm indicadores de insegurança alimentar a nível individual, familiar ou municipal são de difícil acesso, mesmo para órgãos estatais. Desse modo, governos estaduais ou municipais que desejam estabelecer políticas públicas de combate à fome e à insegurança alimentar não encontram amparo em estatísticas e, portanto, têm dificuldade em saber em quais localidades concentrar esforços.

Devido à escassez de dados sobre fome e insegurança alimentar a nível local e municipal, o presente trabalho visa estimar tais indicadores a partir de dados socioeconômicos, que costumam ser mais acessíveis. Para isso, foram usados modelos de aprendizado de máquina (AM).

A metodologia proposta consiste em treinar modelos baseados em árvores de decisão – em especial, modelos de *Boosting* – com dados socioeconômicos extraídos do Cadastro Único para Programas Sociais (CadÚnico)[1], e usá-los para predizer os indica-

[1]Os dados do CadÚnico podem ser encontrados na plataforma de Consulta, Seleção e Extração de

dores de insegurança alimentar Baixo Peso para a Idade (BPI), Baixo Peso para a Altura (BPA) e Baixa Altura para a Idade (BAI), obtidos no Sistema de Vigilância Alimentar e Nutricional (SISVAN)[2]. Os dados serão sumarizados por município, no entanto, acredita-se que a metodologia proposta consiga estimar indicadores de regiões menores (como bairros e povoados), caso os dados do CadÚnico estejam integralmente disponíveis.

Espera-se que o presente estudo mitigue problemas de difícil coleta ou falta de dados, orientando governanças municipais e estaduais que desejem estabelecer políticas públicas de combate à fome baseadas em estatísticas e evidências. Trata-se de uma abordagem inovadora e, até onde sabemos, nunca aplicada a dados nacionais.

2. Trabalhos Relacionados

Nesta seção, é apresentado um breve apanhado de estudos que utilizam AM para predizer algum indicador de insegurança alimentar. Em geral, modelos de árvore se destacaram tanto pelo seu desempenho superior, quanto pela sua interpretabilidade. Tais fatores ajudam justificar a implementação desses modelos no presente estudo.

Em [Subianto et al. 2023], os autores utilizaram uma variedade de fatores socioeconômicos e demográficos, como renda, educação, acesso a serviços de saúde e nutrição para identificar as variáveis que caracterizam ocorrências de insegurança alimentar nos domicílios da província de Aceh, na Indonésia. Eles aplicaram o algoritmo CatBoost para classificar os domicílios em categorias de segurança alimentar. Para entender melhor como cada variável influencia as classificações, eles utilizam o *Shapley Additive Explanations* (SHAP), técnica que define a contribuição individual de cada variável para as predições do modelo, conferindo maior interpretabilidade aos resultados. Esse processo permitiu que os pesquisadores identificassem as variáveis mais importantes para prever a insegurança alimentar, oferecendo sugestões valiosas para a criação de políticas públicas direcionadas a melhorar a segurança alimentar.

O SHAP também é utilizado como parte de uma técnica denominada Boruta-SHAP, particularmente eficaz em cenários de redução de dimensionalidade. Em [Mindiyarti et al. 2023], por exemplo, essa abordagem reduziu um conjunto de 24 indicadores socioeconômicos relacionados à insegurança alimentar, a somente sete variáveis essenciais. Em razão da necessidade de lidar com um banco de dados substancialmente maior, contendo 240 variáveis preditoras, o Boruta-SHAP também foi adotado no presente trabalho para selecionar os atributos mais relevantes.

No estudo [Machefer et al. 2025], os autores avaliaram vários modelos de aprendizado de máquina na predição de insegurança alimentar aguda. Dentre os modelos tesados, XGBoost e Random Forest (RF) apresentaram melhor desempenho, com maior interpretabilidade e robustez. O estudo também enfatizou a necessidade de modelos interpretáveis (utilizando ferramentas como o SHAP, LASSO e coeficientes lineares) para garantir que as previsões do modelo possam ser compreendidas pelos tomadores de decisão.

Alguns estudos brasileiros já demonstraram o potencial de variáveis socioe-

Informações do CadÚnico (CECAD): `https://cecad.cidadania.gov.br/tab_cad.php`.

[2]Os dados do SISVAN podem ser encontrados em: `https://sisaps.saude.gov.br/sisvan/relatoriopublico/index`.

conômicas para estimar a fome ou a insegurança alimentar, porém com escopos ou granularidades restritas. [Barbosa and Nelson 2016] utilizaram *support-vector machines* (SVM) para classificar domicílios rurais do Nordeste como seguros ou inseguros, obtendo acurácia acima de 75%, porém restrita a um recorte regional e binário. No âmbito estadual, [Lobo et al. 2024] integraram diferentes bases de dados e, por meio de classificadores baseados em árvores, previram categorias do Índice Paulista de Vulnerabilidade Social (IPVS), obtendo *F-score* entre 80% e 87%. Em contraste, o trabalho de [Gubert et al. 2010] utilizou regressão logística estimar a prevalência de insegurança alimentar grave em todos os municípios brasileiros com base em PNAD-2004 e Censo-2000.

O presente estudo amplia o estado da arte ao combinar os seguintes fatores: (1) cobertura de quase todos os municípios brasileiros com dados recentes do CadÚnico (setembro de 2023); (2) predição de indicadores contínuos (BPI, BPA, BAI), permitindo monitoramento quantitativo, em vez de classes discretas; e (3) emprego de um *pipeline* com modelos interpretáveis de *boosting* (que figuram o estado da arte no processamento de dados tabulares), aliados ao Boruta-SHAP, para redução de dimensionalidade sem perda de desempenho. Assim, combinam-se granularidade municipal, métricas antropométricas contínuas e modelos sofisticados de AM, de modo a suprimir lacunas de escopo, atualização temporal e interpretabilidade deixadas pelos trabalhos brasileiros citados.

3. Materiais e Métodos

A seguir, os principais passos da metodologia experimental proposta são descritos, incluindo as estratégias de aquisição dos dados e de treinamento e avaliação dos modelos preditivos. A base de dados e os códigos-fonte, usados para implementação da metodologia, estão disponíveis em: `https://github.com/gabriel-s/Hunger_analysis`. No link também é possível encontrar informações complementares que não couberam neste artigo, como a descrição das variáveis retiradas do CECAD.

3.1. Obtenção dos Dados

Foram obtidos, no site do SISVAN, os seguintes indicadores antropométricos para crianças menores de cinco anos: BPI, BPA e BAI. Esses indicadores são referentes a cada município brasileiro e estão em termos percentuais em relação ao número total de crianças analisadas em cada município.

Além disso, foram obtidos, por meio de *web scraping*, os dados do CadÚnico, disponibilizados na plataforma CECAD. Os dados também são referentes a cada município brasileiro (5572 amostras) e estão em termos percentuais em relação ao número total de indivíduos cadastrados no CadÚnico por município.

Vale ressaltar que 38 municípios apresentaram valores ausentes em pelo menos uma das variáveis analisadas e, portanto, foram removidos do conjunto de dados, o qual passou a ter 5534 amostras. As bases do SISVAN e do CECAD foram integradas a partir dos dados de cada município.

3.2. Pré-processamento, Treinamento e Otimização

Optou-se por *ensembles* baseados em árvores (RF e, sobretudo, modelos de *boosting*) por combinarem bom desempenho em dados tabulares, capacidade de lidar com dados

mistos (numéricos e categóricos) e alta interpretabilidade via medidas de importância (e.g., SHAP) [Mienye and Jere 2024]. Além disso, há forte evidência de que modelos de árvores se sobressaem no processamento de dados tabulares, de modo a superar modelos profundos, como o *Multi-Layer Perceptron*, ResNet e sofisticadas arquiteturas baseadas em *Transformers*, como o *FT_Transformer* e o *Self-Attention and Intersample Attention Transformer* (SAINT) [Grinsztajn et al. 2022].

[Bentéjac et al. 2021] compararam diferentes modelos baseados em árvore (RF, *Gradient Boosting* (GB), XGBoost, LightGBM e CatBoost), os quais foram testados em 28 conjuntos de dados diferentes. Aqui, adotou-se um procedimento experimental semelhante ao de Bentéjac *et al.* para determinar o modelo que melhor se encaixa à base de dados utilizada neste trabalho.

Inicialmente, o conjunto de dados foi separado em treino e teste. O conjunto de teste corresponde aos dados de todos os municípios do estado do Ceará (184 amostras) e o conjunto de treinamento corresponde aos dados dos demais municípios (5350 amostras). Essa escolha foi feita para saber se o modelo utilizado seria genérico o suficiente para estimar, precisamente, os indicadores de um estado cujos municípios são todos desconhecidos. Caso o modelo obtenha bom desempenho no conjunto de teste, confirma-se a hipótese de que os indicadores antropométricos podem ser (quase) completamente explicados por meio de indicadores socioeconômicos.

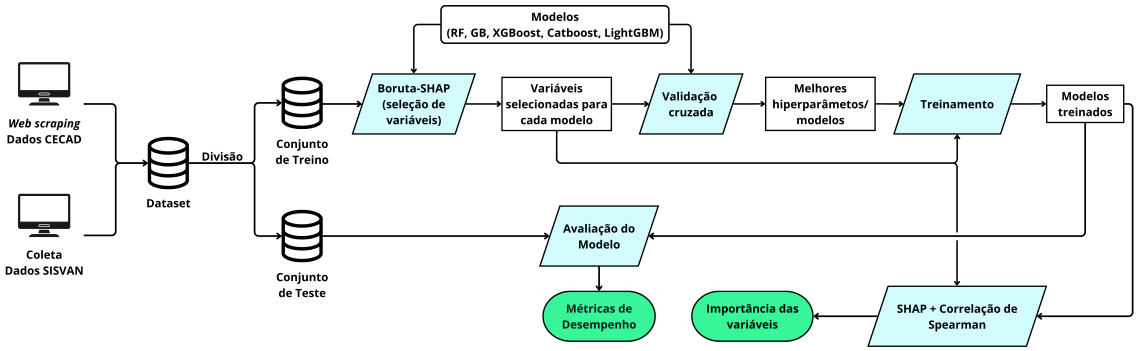


Figura 1. Sumarização das etapas da metodologia proposta.

Todos os modelos foram implementados usando bibliotecas Python[3]. Antes do treinamento, foram selecionadas as variáveis mais relevantes para a predição de cada indicador usando o Boruta-SHAP. O algoritmo usado na seleção de variáveis foi o mesmo utilizado na regressão, isto é, se a regressão for realizada por um modelo RF, as variáveis preditoras também serão selecionadas por um modelo RF.

Cada modelo de seleção de variáveis é composto por 200 árvores e, no caso da RF e do GB, cada árvore tem uma altura máxima igual a 7 (escolha feita para otimizar o tempo de execução). Os demais hiperparâmetros são o padrão das bibliotecas utilizadas.

Em seguida, os hiperparâmetros de cada modelo foram otimizados por meio de uma busca em grade (veja a Tabela 1). Uma validação cruzada (VC) de 10 *folds* no

[3]Para a implementação da RF e GB, do XGBoost, do LightGBM e do CatBoost, usaram-se, respectivamente, as seguintes bibliotecas: scikit-learn [`https://scikit-learn.org`], xgboost(dmlc) [`https://github.com/dmlc/xgboost`], LightGBM(Microsoft) [`https://github.com/microsoft/LightGBM`] e catboost(Yandex) [`https://github.com/catboost/catboost`].

Tabela 1. Hiperparâmetros considerados na busca em grade.

	Hiperparâmetro	Valores da busca em grade
RF	*max_depth*	5; 8; 10
	min_samples_split	2; 5; 10; 20
	min_samples_leaf	1; 25; 50; 70
	max_features	log2; 0,25; sqrt; 1,0
GB	*learning_rate*	0,025; 0,05; 0,1; 0,2; 0,3
	max_depth	2; 3; 5; 7; 10
	max_features	log2; 0,25; sqrt; 1,0
	subsample	0,15; 0,5; 0,75; 1,0
XGBoost	*learning_rate*	0,025; 0,05; 0,1; 0,2; 0,3
	max_depth	2; 3; 5; 7; 10; 100
	colsample_bylevel	0,25; 1,0
	subsample	0,15; 0,5; 0,75; 1,0
LightGBM	*learning_rate*	0,025; 0,05; 0,1; 0,2; 0,3
	num_leaves	3; 7; 15; 31; 127; 1024
	top_rate	0,2; 0,4; 0,6; 0,7
	other_rate	0,05; 0,1; 0,3
	feature_fraction_bynode	log2; 0,25; sqrt; 1,0
CatBoost	*learning_rate*	0,025; 0,05; 0,1; 0,2; 0,3
	max_depth	3; 6; 9
	leaf_estimation_iterations	1; 10
	l2_leaf_reg	1; 3; 6; 9

conjunto de treinamento avaliou o desempenho de cada conjunto de hiperparâmetros com base no *Root Mean Squared Error* (RMSE).

Os modelos (e seus respectivos hiperparâmetros) com melhor desempenho na VC foram, para cada indicador, selecionados e treinados com todo o conjunto de treinamento. Por fim, esses modelos foram avaliados no conjunto de teste. Os passos da metodologia proposta se encontram sumarizados na Figura 1.

4. Resultados

A Tabela 2 mostra o valor médio do RMSE e o desvio padrão (após o sinal $\pm$) da melhor configuração de hiperparâmetros para cada modelo testado por VC. A tabela também contém o número de variáveis definidas como "importante" ou "tentativa" pelo Boruta-SHAP, as quais foram utilizadas por cada modelo em cada cenário preditivo.

Vê-se que o desempenho dos modelos é bem semelhante, estando todos empatados considerando o desvio padrão. Foi escolhido, portanto, para cada indicador, o modelo que utilizou menos variáveis na VC. Tal escolha é feita com base na: (1) eliminação de variáveis ruidosas, que não contribuem para a predição; (2) redução de redundância. Variáveis correlacionadas carregam informação duplicada, logo selecionar um grupo representativo preserva poder preditivo sem inflar a complexidade; e (3) interpretabilidade. Modelos que dependem de um conjunto reduzido de variáveis permitem reportes ob-

jetivos sobre quais fatores socioeconômicos mais influenciam os indicadores de fome, reforçando a confiança na abordagem. Desse modo, escolheu-se o LightGBM para os indicadores BPI e BPA, e o XGBoost para o indicador BAI.

Tabela 2. Desempenho dos modelos na VC.

	BPI		**BPA**		**BAI**	
Modelo	RMSE	Nº var.	RMSE	Nº var.	RMSE	Nº var.
RF	$2,302 \pm 0,498$	16	$2,406 \pm 0,516$	11	$4,880 \pm 0,531$	41
GB	$2,34 \pm 0,486$	20	$2,429 \pm 0,534$	13	$4,9 \pm 0,526$	22
XGBoost	$2,304 \pm 0,488$	18	$2,405 \pm 0,527$	6	$4,875 \pm 0,530$	18
CatBoost	$2,295 \pm 0,477$	16	$2,409 \pm 0,538$	16	$4,888 \pm 0,510$	26
LightGBM	$2,306 \pm 0,482$	8	$2,398 \pm 0,502$	5	$4,888 \pm 0,525$	24

Treinaram-se, portanto, os modelos selecionados com todo o conjunto de treinamento. Os resultados da avaliação desses modelos no conjunto de teste podem ser observados na Tabela 3. Foram calculadas também as métricas *Mean Absolute Error* (MAE) e *Mean Absolute Percentage Error* (MAPE).

Tabela 3. Métricas de desempenho no conjunto de teste dos modelos selecionados por VC.

	RMSE	**MAE**	**MAPE**
BPI (LigthGBM)	1,32	1,1	0,57
BPA (LigthGBM)	2,22	1,45	0,43
BAI (XGBoost)	3,93	2,6	0,22

A Tabela 4 mostra, em termos percentuais, a importância de cada variável preditiva para o processo de regressão, por meio do SHAP. Além disso, a coluna "Correlação" contém o valor da correlação de Spearman entre cada preditor e variável alvo. Por simplicidade, optou-se por mostrar apenas as variáveis com importância maior que 5% e correlação de Spearman maior que $0{,}2$ em valor absoluto.

A Figura 2 mostra os gráficos do erro percentual absoluto de cada município para os indicadores BAI, BPA e BPI.

5. Discussão

Na Tabela 3, nota-se que a BAI tem o maior MAE e o menor MAPE. Isso acontece, porque o indicador de BAI assume, em geral, valores mais elevados em comparação aos indicadores de BPA e BPI. Essa diferença de escala faz com que, mesmo um erro absoluto mais elevado, resulte em um erro percentual menor, já que os valores de referência (valores reais) são elevados.

Como, dentre as métricas utilizadas, o MAPE é a única que permite comparação entre escalas distintas, pode-se dizer que o indicador de BAI (que tem o menor MAPE) é o que melhor se correlaciona com os dados socioeconômicos. De forma análoga, o indicador BPI é o que pior se correlaciona com os dados socioeconômicos. Esse resultado

Tabela 4. Relevância das variáveis preditivas.

	Variável	Importância	Correlação
BAI	Faixa etária / Entre 18 a 24	11,24%	0,418
	Faixa etária / Entre 55 a 59	10,43%	-0,235
	Faixa da renda total da família / Até 1 Salário Mínimo	10,07%	0,42
	Curso que a pessoa frequenta / Ensino Fundamental regular (duração 9 anos)	8,29%	0,23
	Recebe Programa Bolsa Família / Sim	8,28%	0,428
	Cor ou raça / Branca	8,25%	-0,41
	Água canalizada no domicílio / Não	6,79%	0,383
	Faixa etária / Entre 7 a 15	6,65%	0,247
	Material predominante no piso do domicílio / Terra	5,83%	0,374
BPA	Material predominante nas paredes externas do domicílio / Taipa não revestida	21,5%	0,318
	Cor ou raça / Parda	19,58%	0,439
	Cor ou raça / Branca	19,33%	-0,457
BPI	Cor ou raça / Branca	20,17%	-0,38
	Material predominante no piso do domicílio / Terra	16,08%	0,30
	Função principal / Empregado com carteira de trabalho assinada	14,75%	-0,286
	Forma de escoamento sanitário / Vala a céu aberto	10,33%	0,237
	Faixa etária / Entre 16 a 17	9,75%	0,274

já era esperado, pois o BAI, por ser o resultado do acúmulo (ou falha no acúmulo) de crescimento, reflete a desnutrição crônica. O BPI e, principalmente, o BPA, por outro lado, possuem alta instabilidade estatística, pois são resultado de mudanças corporais rápidas e altamente variáveis, respondendo fortemente a infecções, choques de oferta alimentar e fenômenos sazonais, característicos de desnutrição aguda.

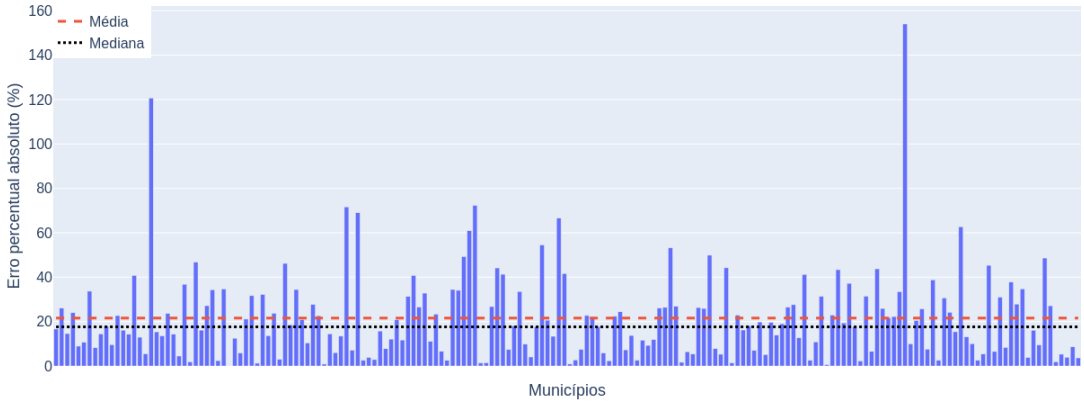
Comparando as Tabelas 2 e 3, vê-se que o RMSE dos modelos no conjunto de teste, que contém somente dados de municípios do Ceará, é próximo (e até menor) que o RMSE obito na VC, feita com municípios dos demais estados. O fato desses valores serem comparáveis indica a capacidade do modelo em generalizar para municípios de localidades não presentes no treinamento.

Na Tabela 4, observa-se que, independentemente da métrica analisada, os preditores relacionados a cor/raça, faixa etária e materiais predominantes no domicílio estão sempre entre os mais importantes. Além disso, a partir dos resultados das correlações de Spearman, é possível chegar a algumas conclusões, detalhadas a seguir.

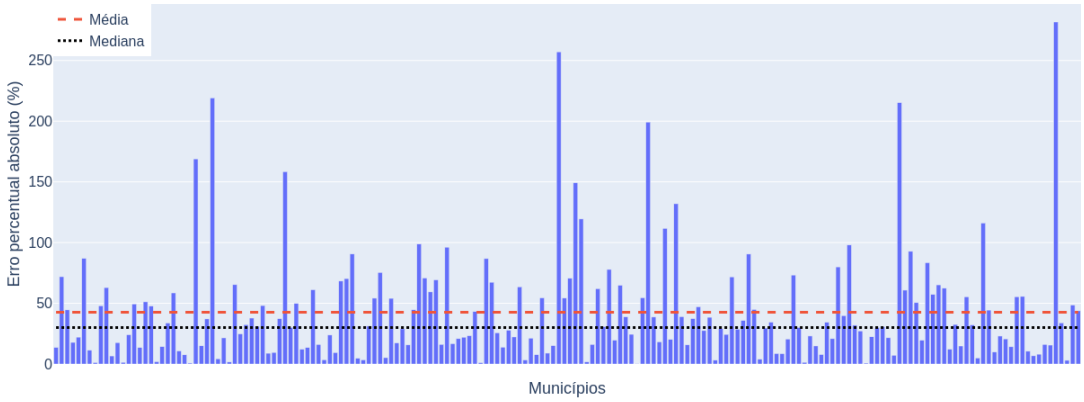
As correlações entre os indicadores antropométricos e a cor/raça branca são sempre negativas e relativamente altas (em valor absoluto). Isso significa que, quanto menor a população branca da região, maiores tendem a ser os índices de fome e insegurança alimentar, ressaltando o viés racial na desigualdade de acesso à alimentação de qualidade.

As correlações entre os indicadores antropométricos e as variáveis relacionadas aos materiais predominantes no domicílio tendem a ser positivas quando tais materiais são precários, como taipa e terra. Isso é um indicativo de que, quanto maior a quantidade de residências precárias em um município, maiores os indicadores de fome.

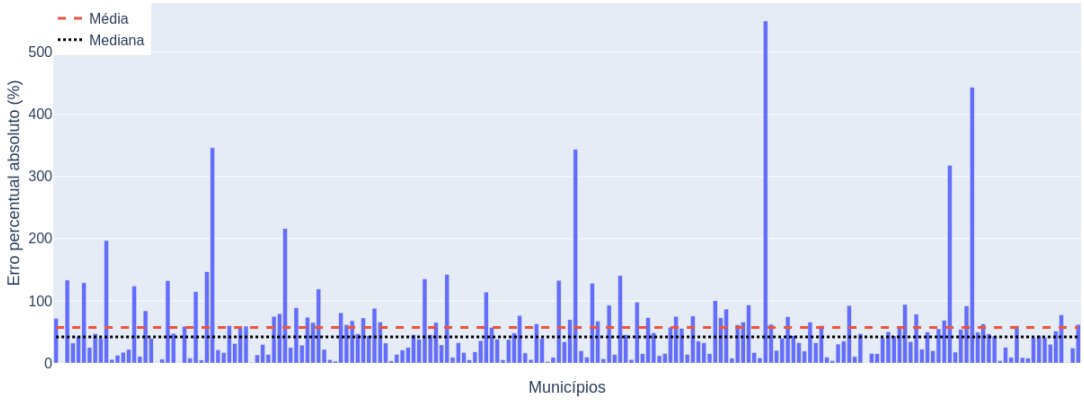
As correlações entre os indicadores antropométricos e a faixa etária tendem a ser



(a) BAI



(b) BPA



(c) BPI

Figura 2. Erro percentual absoluto para cada município do conjunto de teste.

positivas quando se tratam de pessoas jovens (até 24 anos) e negativas quando se tratam de pessoas mais velhas (entre 55 e 59 anos). Esse comportamento pode ser um reflexo da expectativa de vida, já que, em locais com elevadas expectativas de vida, a população tende a ter melhor acesso à nutrição e serviços de saúde.

A Figura 2 apresenta, para cada município cearense, o erro percentual absoluto (APE) obtido pelos modelos nas três tarefas de predição. Sobre cada distribuição, foram desenhados dois marcadores: a linha vermelha representa a mediana do APE, enquanto a linha preta indica a média do APE. A mediana é pouco sensível a grandes *outliers* e fornece uma noção do erro típico na metade dos municípios. Já a média reflete o erro global, e é puxada para cima pelos poucos municípios com erros extremamente altos. Assim, o fato de a média ser maior que a mediana em todos os indicadores, evidencia a existência de registros ruidosos nos dados do SISVAN, os quais inflam a média sem afetar a mediana.

Na Figura 2, nota-se também uma distribuição de erros consideravelmente assimétrica. No indicador BAI, por exemplo, 75% dos municípios exibem erro percentual absoluto (APE) abaixo de 30%, mas uma pequena fração ultrapassa 100%. Para o BPI, a situação é ainda mais extrema: há municípios cujo APE atinge 550%, patamar claramente incompatível com a prevalência real desses distúrbios nutricionais. Esses picos são esporádicos, não se concentram nos mesmos municípios, nem se replicam entre os três indicadores, o que sugere falhas pontuais na coleta ou no registro dos dados. Essa interpretação encontra ainda mais respaldo em [Silva et al. 2023], onde foram analisados mais de 70 milhões de registros do SISVAN (2008–2017) e constatados desvios-padrão dos escores-z entre 1,2 e 1,6, o que está acima do valor de referência da OMS ($\approx 1,0$), indicando medições imprecisas, arredondamento de idade ou cobertura amostral insuficiente.

Convém ressaltar que o SISVAN não acompanha toda a população infantil brasileira. Em vez disso, são acompanhados apenas usuários dos serviços de Atenção Primária à Saúde, em sua maioria beneficiários do Programa Bolsa Família, isto é, o segmento socioeconômico mais vulnerável [Silva et al. 2023, Mrejen et al. 2023]. Mesmo assim, a cobertura permaneceu limitada: a proporção de crianças menores de cinco anos com registros antropométricos subiu de 17,7% em 2008 para apenas 45,4% em 2017, ainda menos da metade do público-alvo [Silva et al. 2023]. Portanto, a cobertura do SISVAN ainda é incipiente na maioria das regiões e unidades federativas do país [Silva et al. 2023].

Há lacunas evidentes entre o número de crianças considerada em risco (por exemplo, aquelas vinculadas ao Programa Bolsa Família) e o universo efetivamente monitorado pelo SISVAN. Esse descompasso evidencia não apenas a seletividade do sistema, mas também fragilidades no monitoramento contínuo. Idealmente, as avaliações nutricionais deveriam ser realizadas sobre amostras aleatória de todas as crianças da localidade, tornando os dados mais representativos e, consequentemente, as análises e predições deste estudo mais próximas da realidade.

6. Conclusão

A fome e a insegurança alimentar no Brasil são difíceis de monitorar localmente devido à limitação de dados diretos, embora haja ampla disponibilidade de informações socioeconômicas no CadÚnico. Este estudo propôs uma abordagem de aprendizado de máquina para estimar indicadores antropométricos de insegurança alimentar, combinando dados do CadÚnico e SISVAN para mais de 5.500 municípios.

Modelos baseados em árvores de decisão (LightGBM, XGBoost e outros) apresentaram bom desempenho, com destaque para o MAPE de 22% no BAI. Os resultados

mostram que variáveis socioeconômicas, como raça/cor, faixa etária e condições habitacionais, têm forte relação com os indicadores, evidenciando o potencial da metodologia para apoiar políticas públicas em regiões com dados nutricionais limitados.

As principais limitações decorrem da baixa cobertura e viés amostral do SISVAN, o que afeta a representatividade e a precisão das estimativas. A aplicação em escalas menores que o município depende do acesso a dados do CadÚnico desagregados por indivíduo ou família.

Conclui-se que a metodologia é viável e transferível para diferentes localidades. Estudos futuros devem explorar dados submunicipais e séries temporais para monitoramento dinâmico da insegurança alimentar.

Agradecimentos

Danielo G. Gomes agradece ao CNPq pela bolsa de produtividade (processo 311845/2022-3) e à FUNCAP pelo apoio financeiro na execução do projeto Cientista Chefe da Transformação Digital do Estado do Ceará (processo: 31052.000465/2025-45).

Referências

Aguiar, I. W. O., Carioca, A. A. F., Barbosa, B. B., Adriano, L. S., Barros, A. Q. S., Kendall, C., and Kerr, L. R. F. S. (2023). Indicadores antropométricos em povos e comunidades tradicionais do Brasil: Análise de registros individuais do sistema de vigilância alimentar e nutricional, 2019. *Epidemiologia e Serviços de Saúde*, 32:e2023543.

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.

Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54:1937–1967.

Food and Agriculture Organization of the United Nations (2023). Putting a number on hunger – interactive presentation. Interactive web page, The State of Food Security and Nutrition in the World 2023. acessado em 05 agosto 2025.

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.

Gubert, M. B., Benicio, M. H. D., and Monteiro, C. A. (2010). Estimativas de insegurança alimentar grave nos municípios brasileiros. *Cadernos de Saúde Pública*, 26(8):1595–1605.

Lobo, P. L. S., Santos, A. C., and Oliveira, M. R. (2024). Aplicação de algoritmos de aprendizado de máquina na análise da vulnerabilidade social e insegurança alimentar. In *Anais do ERI-GO 2024*, pages 158–167. Sociedade Brasileira de Computação.

Machefer, M., Thomas, A.-C., Meroni, M., Pena, J. M. V. L., Ronco, M., Corbane, C., and Rembold, F. (2025). Potential and limitations of machine learning modeling for forecasting acute food insecurity. *Global Food Security*, 45:100859.

Mienye, I. D. and Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE access*.

Mindiyarti, N., Sartono, B., Indahwati, I., Hadi, A. F., and Ramadhani, E. (2023). A study in determining indicators of food-insecure households using SHAP and Boruta SHAP. In *AIP Conference Proceedings*, volume 2720. AIP Publishing.

Morais, D. d. C., Dutra, L. V., Franceschini, S. d. C. C., and Priore, S. E. (2014). Insegurança alimentar e indicadores antropométricos, dietéticos e sociais em estudos brasileiros: uma revisão sistemática. *Ciência & Saúde Coletiva*, 19:1475–1488.

Mrejen, M., Cruz, M. V., and Rosa, L. (2023). O sistema de vigilância alimentar e nutricional (sisvan) como ferramenta de monitoramento do estado nutricional de crianças e adolescentes no brasil. *Cadernos de Saúde Pública*, 39:e00169622.

Silva, N. d. J., Carrilho, T. R. B., Pinto, E. d. J., Andrade, R. d. C. S. d., Silva, S. A., Pedroso, J., Spaniol, A. M., Bortolini, G. A., Fagundes, A., Nilson, E. A. F., et al. (2023). Quality of child anthropometric data from sisvan, brazil, 2008-2017. *Revista de Saúde Pública*, 57:62.

Sousa, I. M. L. d. and Diniz, R. B. (2024). Controle da qualidade e segurança alimentar durante a pandemia por covid-19 nos setores públicos do Brasil. *Nutrivisa - Revista de Nutrição e Vigilância em Saúde*, 11(1):e12302.

Subianto, M., ULYA, I. Y., RAMADHANI, E., SARTONO, B., and HADI, A. F. (2023). Application of SHAP on CatBoost classification for identification of variabels characterizing food insecurity occurrences in aceh province households. *Jurnal Natural*, 23(3):230–244.

World Health Organization (2023). *The State of Food Security and Nutrition in the World 2023: Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum*, volume 2023. Food & Agriculture Org.