Predição da taxa de internação por desnutrição no Sistema Único de Saúde utilizando aprendizado de máquina

Vanessa P. Resmini[1], **Mariana Recamonde-Mendoza**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS), Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

vanessaresmini22@hotmail.com,mrmendoza@inf.ufrgs.br

Abstract. *According to the 2025 Hunger Map, 35 million people in Brazil face food insecurity, demonstrating the need to maintain and structure policies to reduce hunger and poverty. In this study, machine learning models were developed to predict the annual rate of hospitalizations due to malnutrition in the brazilian Sistema Único de Saúde (SUS), segmented by the country's 439 health regions, based on patients place of residence. In the out-of-time dataset, R^2 values of 0.66 were observed for LGBM, 0.64 for Random Forest, and 0.65 for XGBoost, while MAPE values were 0.35, 0.42, and 0.37, respectively. The results also showed variations in model performance among the Brazilian macro-regions.*

Resumo. *De acordo com o Mapa da Fome de 2025, há 35 milhões de pessoas com dificuldade para se alimentar no Brasil, demonstrando a necessidade em manter e estruturar as políticas de redução de fome e pobreza. Neste trabalho, foram desenvolvidos modelos de aprendizado de máquina com o objetivo de predizer a taxa anual de internação hospitalar por desnutrição no Sistema Único de Saúde (SUS), segmentada pelas 439 regiões de saúde do país, com base no local de residência dos pacientes. No conjunto out-of-time, observou-se valores de R^2 de 0,66 para LGBM, 0,64 para RF e 0,65 para XGBoost, enquanto os valores de MAPE foram de 0,35, 0,42 e 0,37, respectivamente. Os resultados também evidenciaram variações no desempenho dos modelos entre as macrorregiões brasileiras.*

1. Introdução

A desnutrição constitui um importante indicador de vulnerabilidade social e sanitária, estando associada não apenas à insegurança alimentar, definida como a limitação no acesso regular e permanente a alimentos em quantidade e em qualidade adequadas, mas também às desigualdades socioeconômicas e territoriais [Valadares et al. 2023]. No Brasil, embora avanços tenham sido registrados na última década, a desnutrição ainda persiste em diversas regiões, especialmente entre populações em situação de pobreza extrema. Mesmo com o Brasil fora do Mapa da Fome, estimativas recentes indicam que cerca de 35 milhões de pessoas enfrentam dificuldades para acessar alimentação adequada [FAO et al. 2025], evidenciando a necessidade de monitoramento contínuo e de políticas públicas eficazes. Entretanto, compreender e antecipar os impactos dessas desigualdades em indicadores de saúde pública ainda representa um desafio para gestores e pesquisadores.

Embora a insegurança alimentar seja um fenômeno influenciado por múltiplos fatores, a taxa de internação hospitalar por desnutrição no Sistema Único de Saúde (SUS)

representa um indicador relevante para compreender os impactos da insegurança alimentar e das desigualdades regionais no Brasil. Com o avanço das técnicas de análise de dados, especialmente de aprendizado de máquina (AM), torna-se possível explorar grandes bases de dados públicas e identificar padrões complexos entre variáveis socioeconômicas, educacionais, demográficas e de infraestrutura associados a fenômenos de saúde pública. Modelos de AM têm sido aplicados em diferentes escalas, incluindo a classificação de insegurança alimentar em nível domiciliar, a análise de fatores associados em contextos regionais e a previsão de insegurança alimentar em cenários de crise [da Silva et al. 2025, Barbosa and Nelson 2016, Cao et al. 2025]. Revisões recentes destacam o potencial dessas abordagens para apoiar sistemas de monitoramento e alerta precoce de insegurança alimentar, embora persistam desafios relacionados à disponibilidade e integração de dados [Jarray et al. 2023]. Apesar desses avanços, ainda são escassos os estudos que investigam modelos preditivos pare desfechos de saúde associados à insegurança alimentar, como internações relacionadas à desnutrição, em escala nacional.

Diante disso, este trabalho propõe o desenvolvimento de modelos de AM para predizer a taxa anual de internações hospitalares por desnutrição no SUS nas 439 regiões de saúde do Brasil, considerando o local de residência dos pacientes. Além da modelagem preditiva, são empregadas técnicas de explicabilidade para identificar os principais fatores associados às predições. Adicionalmente, é realizada uma análise comparativa do desempenho preditivo entre as diferentes macrorregiões do país. Os resultados podem contribuir para o monitoramento e planejamento em saúde pública, auxiliando na identificação de regiões mais vulneráveis e no direcionamento de políticas voltadas à segurança alimentar.

2. Trabalhos relacionados

Estudos demonstram que a insegurança alimentar pode ser prevista com elevada acurácia por meio de variáveis socioeconômicas, mesmo na ausência de dados nutricionais diretos [da Silva et al. 2025]. O estudo de de Silva *et al.* foi realizado com famílias do programa Cartão Mais Infância Ceará e evidenciou que fatores como renda, acesso à água, vínculos formais de trabalho e cobertura de programas sociais estão fortemente associados à insegurança alimentar e, por extensão, à desnutrição. De forma semelhante, aplicações de *Support Vector Machines* demonstraram potencial para classificar domicílios em situação de insegurança alimentar e identificar fatores socioeconômicos relevantes [Barbosa and Nelson 2016]. O estudo mostra que determinantes da insegurança alimentar variam conforme o contexto local. No caso analisado (região agrícola do Nordeste brasileiro), fatores como pobreza estrutural, dependência da agricultura e exposição a secas desempenham papel importante na vulnerabilidade alimentar.

Além de análises em nível domiciliar, pesquisas recentes têm aplicado técnicas de AM para previsão de insegurança alimentar em diferentes escalas geográficas. Estudos internacionais utilizaram dados socioeconômicos, demográficos e ambientais para prever níveis de insegurança alimentar em contextos de crise, demonstrando que algoritmos baseados em árvores podem apresentar bom desempenho na identificação de populações vulneráveis [Cao et al. 2025]. Outra linha de pesquisa utiliza AM para avaliar padrões globais de segurança alimentar, empregando modelos de AM para lidar com limitações de dados, como a imputação de indicadores ausentes e a construção de bases integradas para avaliação comparativa entre países, permitindo identificar disparidades regionais e apoiar a formulação de políticas públicas em escala internacional [Xiong et al. 2024].

Apesar do avanço dessas abordagens, a maior parte da literatura concentra-se na previsão direta de indicadores de insegurança alimentar ou em sua classificação em nível domiciliar ou nacional. Ainda são escassos os estudos que investigam desfechos de saúde associados à insegurança alimentar, como hospitalizações relacionadas à desnutrição, particularmente em escala subnacional. Nesse sentido, este trabalho busca contribuir para a literatura ao explorar a predição da taxa de internações hospitalares por desnutrição no SUS em nível de regiões de saúde, combinando modelagem preditiva e análise de explicabilidade para compreender os fatores associados às variações regionais desse indicador.

3. Metodologia

3.1. Fonte dos dados e definição de atributos

Para este trabalho, o *dataset* foi construído a partir de dados abertos provenientes de diferentes órgãos e autarquias governamentais, abrangendo identificação geográfica e indicadores econômicos, demográficos, sociais, educacionais e de infraestrutura. O resultado é disponibilizado em repositório público[1]. O atributo-alvo deste estudo é a taxa anual de internações hospitalares por desnutrição no SUS, considerando o local de residência dos pacientes. Esse indicador foi obtido a partir do DATASUS e calculado como o número de internações por 100.000 habitantes. Os dados foram agregados nas 439 regiões de saúde do SUS, que correspondem a agrupamentos de municípios vizinhos definidos pelo Ministério da Saúde com o objetivo de integrar a organização, o planejamento e a execução de ações e serviços de saúde.

Os atributos explicativos abrangem o período de 2012 a 2021, enquanto o atributo-alvo foi coletado entre 2013 e 2022. O ano de 2022 foi reservado para validação *out-of-time* (OOT) dos modelos. A Tabela 1 abrange os atributos a serem analisados e suas origens. Todos os dados são intrinsecamente numéricos e fornecidos na granularidade de municipalidade, ou seja, em 2012 seriam 5.568 instâncias; por sua vez, estas instâncias são agrupadas nas 439 regiões do SUS através de soma. Exceções desta estratégia são o salário médio de admissão, sobre o qual é realizada uma média de todos os salários formais de admissão, utilizando dezembro como o mês de referência, assim como as estatísticas oficiais, e do percentual da população que reside em município com Política de Saneamento Básica.

Para reduzir possíveis efeitos de variações anuais e capturar relações temporais entre variáveis explicativas e o desfecho de interesse, os atributos explicativos foram utilizados na forma de médias móveis dos dois anos anteriores a cada instante t. Assim, para prever o atributo-alvo no ano $t + 1$, utiliza-se a média dos atributos explicativos observados em t e $t - 1$. Em função desse procedimento, o período efetivamente utilizado para modelagem do atributo-alvo compreende os anos de 2014 a 2022, sendo 2022 reservado para avaliação *OOT*.

3.2. Pré-processamento dos dados

O conjunto de dados foi padronizado utilizando o método *StandardScaler* da biblioteca *scikit-learn*. Todos os atributos são padronizados, com exceção do percentual de mulheres na população, percentual de adesão à política de saneamento e o valor médio de admissão

[1] https://doi.org/10.5281/zenodo.19870509

do salário formal, pois já são valores normalizados no caso dos percentuais ou que não variam em larga escala dentro do país, como é o caso do atributo salarial. A padronização foi realizada dentro do subconjunto de treino e depois aplicada para os subconjuntos de testes e validação, visando evitar *data leakage* no desenvolvimento dos modelos.

Tabela 1. Atributos explicativos incluídos no treinamento dos modelos.

Atributo	Descrição	Fonte
vlr_pib_corrente_1000	PIB corrente anual, em milhares de reais	IBGE
vlr_pbf	Valor anual transferido pelo Programa Bolsa Família (PBF) e Auxílio Brasil	SAGICAD
qtd_familias_pbf	Quantidade de famílias atendidas durante o ano pelo PBF e Auxílio Brasil	SAGICAD
vlr_bpc	Valor anual transferido pelo Benefício de Prestação Continuada (BPC)	SAGICAD
qtd_beneficiarios_bpc	Quantidade de beneficiários atendidos durante o ano pelo BPC	SAGICAD
vlr_medio_salario_formal	Salário médio de admissão dos vínculos formais no mês de dezembro	CAGED
qtd_vinculos_formais	Estoque de vínculos empregatícios formais registrados no mês de dezembro	RAIS
qtd_supermercados	Quantidade média de supermercados durante o ano	Receita Federal
perc_mulheres	Percentual de mulheres na população estimada	DataSUS
qtd_populacao	População estimada	DataSUS
qtd_escolas	Quantidade de escolas de ensino básico públicas e particulares	Dados GOV
qtd_alunos_ensino_basico	Quantidade de alunos matriculados no ensino básico	Dados GOV
qtd_alunos_pnae	Quantidade de alunos atendidos pelo Programa Nacional de Alimentação Escolar (PNAE)	Dados GOV
qtd_animais	Quantidade de animais produzidos durante o ano (galináceos, suínos e bovinos)	IBGE
vlr_lavoura_1000	Valor da produção agrícola de alimentos, em milhares de reais	IBGE
vlr_produzida_lavoura_ton	Produção agrícola de alimentos, em toneladas	IBGE
pop_total_atendida_agua	População com acesso à água potável	SNIS
pop_total_atendida_esgoto	População com acesso à rede de esgoto	SNIS
pop_total_atendida_coleta	População com acesso à coleta de resíduos sólidos	SNIS
adesao_politica_saneamento	Percentual da população em município com Política de Saneamento Básica	SNIS

Foi realizada a estratificação dos dados em conjuntos de treino, teste e OOT. Para a validação OOT, foi escolhido sempre o último ano do conjunto. Já os anos anteriores foram estratificados entre treino e teste, com a proporção de 70% das Regiões do SUS, dentro de cada ano, para treino e 30% para teste.

3.3. Treinamento e otimização de modelos e seleção de atributos

Os modelos de regressão foram treinados com os algoritmos *Light Gradient Boosting Machine* (LGBM), *Random Forests* (*RF*) e *eXtreme Gradient Boosting* (*XGBoost*), considerando-se os dados para todas as regiões de saúde analisadas (*i.e.*, modelo global). Para a identificação dos atributos mais relevantes, foi aplicada a técnica de seleção de atributos *recursive feature elimination* (RFE) sobre os dados normalizados no conjunto

de treino, utilizando o mesmo algoritmo adotado no treinamento do modelo principal (*i.e.*, regressor). Este processo foi baseado na avaliação do R^2 com validação cruzada de cinco *folds* e passo unitário, ou seja, a cada iteração, um atributo é removido. O número de atributos escolhidos é aquele que maximiza o desempenho médio da validação cruzada.

A seleção dos melhores hiperparâmetros para os modelos foi feita com otimização bayesiana. Neste trabalho, optou-se por trinta e duas iterações, pois um número maior de iterações não resultava em melhora significativa da métrica Para o trabalho, definiu-se como objetivo do otimizador a maximização do R^2 com validação cruzada de cinco *folds*.

Adicionalmente, para investigar possíveis diferenças regionais, foram treinados modelos específicos para cada macrorregião do país, utilizando apenas os dados correspondentes a cada macrorregião. Nesses casos, os mesmos procedimentos de construção das variáveis, seleção de atributos e busca de hiperparâmetros foram aplicados de forma independente, com o objetivo de capturar eventuais heterogeneidades regionais. Os resultados desses modelos foram então comparados aos obtidos pelos modelos globais.

Por fim, para estabelecer uma referência (*baseline*) para comparação dos modelos propostos, também foi treinado um modelo de regressão linear com as mesmas aplicações de *feature engineering*. Esse modelo permite avaliar se os algoritmos de AM empregados apresentam ganhos de desempenho em relação a uma abordagem linear simples.

3.4. Avaliação de desempenho e explicabilidade de modelos

O desempenho dos modelos foi avaliado com base em duas métricas. A primeira foi o coeficiente de determinação (R^2), seguindo outros trabalhos de cunho semelhante [Almalki et al. 2021, Almeida et al. 2024]. O R^2 é bastante popular, porém recebe críticas de alguns autores visto que esta métrica olha apenas o comportamento das variáveis em relação uma a outra, sem considerar seus valores [Li 2017]. Isso significa que um modelo pode ter um R^2 elevado e ainda assim estar errando mais do que o esperado nos valores previstos, logo, esta é uma métrica que deve ser acompanhada de outras métricas baseadas em residuais. Desta forma, também foi avaliado o *mean absolute percentage error* (MAPE), que calcula um erro percentual, tornando mais fácil o comparativo entre diferentes modelos.

Para a explicabilidade dos modelos, foi utilizado o SHAP, pois permite identificar os atributos mais relevantes para cada modelo, além de compreender a direção do impacto, se positivo ou negativo, de cada variável sobre a taxa prevista. Além disso, como se busca entender as diferenças e semelhanças de cada região do país, o SHAP auxilia na compreensão do comportamento de cada atributo dentro de cada região.

3.5. Redução de dimensionalidade para visualização de padrões

Foram aplicadas técnicas de redução de dimensionalidade com o objetivo de investigar possíveis diferenças no comportamento dos atributos entre as macrorregiões do Brasil e verificar se esses atributos permitem distinguir as regiões. Para isso, foi utilizada a projeção t-SNE, uma técnica não linear de redução de dimensionalidade amplamente empregada para visualização de dados de alta dimensão. O método busca preservar as relações de proximidade entre os pontos no espaço original ao projetá-los em um espaço de menor dimensão, priorizando a preservação de pequenas distâncias entre pares de observações [Awan 2024].

4. Resultados

4.1. Atributos selecionados para os modelos globais

Após a padronização dos atributos, foi realizada a seleção de atributos com a técnica RFE para cada modelo. A métrica R^2 melhora conforme aumenta o número de atributos e depois passa a variar muito pouco. Assim, chegou-se à conclusão de que não é necessário se utilizar de todos os vinte atributos, podendo-se atingir resultados semelhantes, ou até superiores, com menos atributos. Os atributos explicativos selecionados para cada modelo são apresentados na Tabela 2.

Tabela 2. Atributos selecionados para cada modelo.

Atributo	*LGBM*	*RF*	*XGBoost*	Regressão linear
vlr_pib_corrente_1000_media_2anos	x	x	x	x
vlr_pbf_media_2anos				x
qtd_familias_pbf_media_2anos	x	x	x	x
vlr_bpc_media_2anos				x
qtd_beneficiarios_bpc_media_2anos	x	x	x	x
vlr_medio_salario_formal_media_2anos				x
qtd_vinculos_formais_media_2anos	x	x	x	x
qtd_supermercados_media_2anos			x	x
perc_mulheres_media_2anos	x	x	x	x
qtd_populacao_media_2anos	x	x	x	x
qtd_escolas_media_2anos	x	x	x	x
qtd_alunos_ensino_basico_media_2anos		x	x	x
qtd_alunos_pnae_media_2anos	x		x	x
qtd_animais_media_2anos	x	x	x	x
vlr_lavoura_1000_media_2anos	x	x	x	x
qtd_produzida_lavoura_ton_media_2anos	x	x	x	x
pop_total_atendida_agua_media_2anos				x
pop_total_atendida_esgoto_media_2anos	x	x	x	x
pop_total_atendida_coleta_media_2anos				x
adesao_politica_saneamento_media_2anos				x

Para os modelos *LGBM* e *RF*, o melhor número de atributos é 12, enquanto para o *XGBoost*, o melhor número de atributos é 14 e para a regressão linear é de 20. Os resultados são resumidos na Tabela 2, onde podemos observar, excetuando o *baseline*, que a maioria dos atributos se repete para os três modelos; por outro lado, alguns atributos não foram selecionados para nenhum modelo (*i.e.*, vlr_pbf, vlr_bpc, vlr_medio_salario_formal, pop_total_atendida_agua, pop_total_atendida_coleta e adesao_politica_saneamento). É interessante notar que, embora os atributos relacionados ao valor dos programas de repasse de renda não foram selecionados, os atributos que indicam a quantidade de pessoas atendidas por estes programas se mostraram relevantes. Também, mostra-se importante o atributo qtd_alunos_pnae, que indica quantos estudantes são atendidos pelo PNAE, uma iniciativa voltada a garantir uma alimentação nutritiva e equilibrada para alunos que, em muitos casos, têm na escola sua única refeição completa do dia [Simeon 2023]. Dentre os atributos selecionados por todos os modelos, estão vlr_pib_corrente_1000, qtd_familias_pbf e qtd_vinculos_formais, reforçando que contribuem de forma significativa para a predição da taxa de internação por desnutrição nos dados analisados.

4.2. Análise de desempenho para os modelos globais

A Tabela 3 sumariza o desempenho dos modelos globais, ou seja, considerando todas as regiões de saúde, para cada algoritmo utilizado. Para os modelos de AM propostos, os resultados se referem aos modelos após a seleção de atributos e a otimização bayesiana. Os resultados indicam que os modelos baseados em aprendizado de máquina apresentam desempenho substancialmente superior ao modelo de regressão linear utilizado como *baseline*. Enquanto os modelos *LGBM*, *RF* e *XGBoost* alcançam valores de R^2 elevados nos conjuntos de teste e OOT (entre 0,64 e 0,85), o modelo linear apresenta coeficientes de determinação próximos de zero, indicando baixa capacidade explicativa. De forma consistente, os valores de MAPE também são significativamente menores nos modelos baseados em árvores, evidenciando maior precisão preditiva. Esses resultados indicam que uma abordagem linear simples não é capaz de capturar adequadamente a relação entre os atributos explicativos e a taxa de internação por desnutrição, enquanto modelos mais complexos consegues aprender, ao menos parcialmente, estes padrões.

Tabela 3. Desempenho dos modelos globais para os modelos propostos e para o *baseline* baseado em regressão linear.

Métrica	**R^2**			
Conjunto/Modelo	***LGBM***	***RF***	***XGBoost***	**Regressão Linear**
Treino	0,99	0,98	0,99	0,09
Teste	0,83	0,80	0,85	0,08
OOT	0,66	0,64	0,65	0,02
Métrica	**MAPE**			
Conjunto/Modelo	***LGBM***	***RF***	***XGBoost***	**Regressão Linear**
Treino	0,06	0,09	0,01	0,76
Teste	0,22	0,35	0,21	0,81
OOT	0,35	0,42	0,37	0,63

O teste t pareado foi utilizado para comparar os modelos entre si e contra o *baseline*. Considerando a métrica MAPE do OOT, todos os modelos mostraram-se significativamente distintos do *baseline* ($p < 0,001$). Entre os algoritmos, as diferenças foram significativas para os pares *LGBM*/*RF* e *RF*/*XGBoost* ($p < 0,001$) e para o par *LGBM*/*XGBoost* ($p < 0,005$).

Segregou-se as métricas por cada região do país, com o objetivo de se identificar o desempenho regional e se havia alguma diferença. Os modelos foram treinados com dados de todo o Brasil, o mérito desta análise é de observar a diferença de desempenho nas macrorregiões ao não se ter um modelo local. Os resultados das avaliações para as macrorregiões estão na Tabela 4. De modo geral, observa-se que o desempenho varia de forma significativa entre as macrorregiões, tanto em termos de R^2 quanto de MAPE, indicando que a relação entre os atributos explicativos e a taxa de internação por desnutrição não é homogênea no território nacional.

Considerando especificamente o comportamento observado para a região Centro-Oeste, os resultados apresentados na Tabela 4 indicam valores de R^2 muito baixos, próximos de zero, enquanto o MAPE se mantém em um patamar intermediário. Diante de um R^2 tão reduzido, seria esperado um valor de MAPE mais elevado, próximo de um; contudo, esse comportamento não se verifica em função da distribuição dos resíduos observada nessa região.

Tabela 4. Métricas dos modelos globais abertas por macrorregião.

Modelo	*LGBM*		*RF*		*XGBoost*	
Macrorregião/Métrica	**R^2**	**MAPE**	**R^2**	**MAPE**	**R^2**	**MAPE**
Centro-oeste	-0,0005	0,54	-0,09	0,64	-0,05	0,55
Nordeste	0,44	0,35	0,17	0,48	0,33	0,41
Norte	0,29	0,32	0,39	0,34	0,37	0,33
Sudeste	0,72	0,34	0,72	0,38	0,72	0,34
Sul	0,48	0,25	0,43	0,28	0,45	0,25
Brasil	0,66	0,35	0,64	0,42	0,65	0,37

O Centro-Oeste é a região com o menor número de regiões de saúde, o que impõe limitações adicionais à construção de um modelo global para todo o Brasil, em razão do desbalanceamento na quantidade de dados disponíveis para treinamento em comparação às demais macrorregiões. Como consequência, os modelos não conseguem explicar de forma satisfatória a variabilidade do atributo-alvo nessa macrorregião, resultando em valores de R^2 muito baixos, conforme evidenciado na Figura 1. Ainda assim, observa-se que os erros percentuais médios não são extremos, o que justifica a manutenção do MAPE em níveis intermediários, apesar da baixa capacidade explicativa dos modelos. É possível também, que, estes dados mostrem uma diferença estrutural do Centro-Oeste em relação às outras macrorregiões, possivelmente associada ao seu processo histórico de formação.

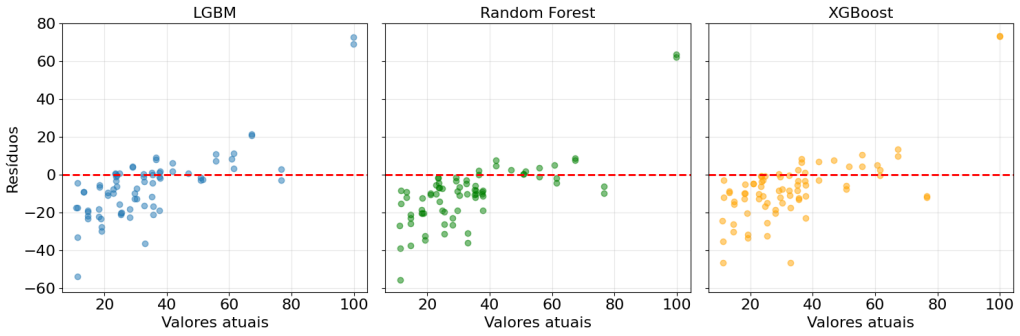


Figura 1. Resíduos vs. valores atuais no conjunto OOT para Centro-Oeste

Observando-se as Figuras 2 e 3, fica evidente que no Sudeste, onde o R^2 é menor, os resíduos são mais concentrados ao redor de zero, enquanto no Centro-Oeste, onde o R^2 é próximo de zero, os resíduos possuem a sua média centrada mais à esquerda, mostrando uma dispersão maior de resíduos. Já no Sul, conforme a Figura 4, há uma certa dispersão dos resíduos, porém o valor absoluto dos resíduos são menores o que leva a um MAPE menor. O Sudeste concentra os melhores resultados entre todas as macrorregiões analisadas. Os três modelos atingem valores elevados e idênticos de R^2 (0,72), indicando forte capacidade explicativa, com MAPE relativamente baixos e semelhantes. Esse comportamento sugere que, nessa macrorregião, os atributos utilizados são particularmente informativos e que o padrão das internações por desnutrição é mais estável e previsível. Isso possivelmente ocorre pelo fato da macrorregião Sudeste apresentar a maior quantidade de regiões de saúde, gerando um desbalanceamento favorável a si. Além disso, os dois estados com mais municípios do Brasil, Minas Gerais e São Paulo, estão na região Sudeste, assim há mais informação para compor os dados das regiões de saúde.

Escolheu-se o modelo *LGBM*, que apresentava as melhores métricas para o Brasil,

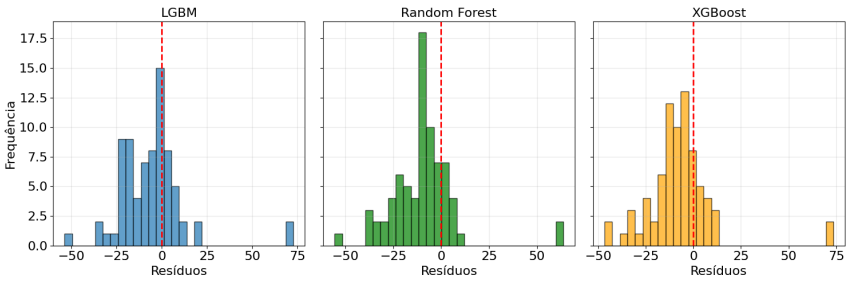


Figura 2. Histograma de resíduos no conjunto OOT para Centro-Oeste

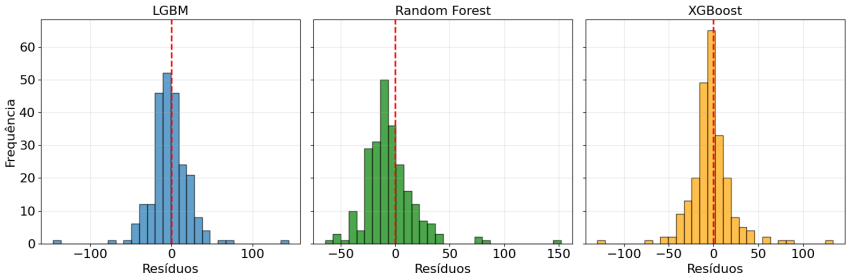


Figura 3. Histograma de resíduos no conjunto OOT para Sudeste

para se analisar o gráfico SHAP, apresentado na Figura 5. Observa-se que o PIB é o atributo de maior influência no modelo. Valores mais elevados desse atributo tendem a estar associados a impactos negativos no valor predito, indicando redução da taxa de internação por desnutrição. Esse comportamento está alinhado com a interpretação socioeconômica do fenômeno, uma vez que maior atividade econômica tende a estar associada a melhores condições de acesso à alimentação e aos serviços de saúde.

Tal comportamento aparece também no atributo da quantidade de vínculos formais, que é o quarto mais importante para o Brasil. Conforme a série histórica da Pesquisa Nacional por Amostra de Domicílios Contínua, fornecida pelo IBGE, trabalhadores formais têm rendimento mensal superior à trabalhadores informais.Chama a atenção que o segundo atributo mais importante é a população com atendimento para coleta de esgoto, embora, de uma maneira anti-intuitiva, quanto maior este atributo, maior a taxa de internação. Assim, esse resultado sugere que o indicador pode estar capturando outros efeitos estruturais, uma vez que a macrorregião Sul não possui bons índices de atendimento de esgoto, por exemplo, ela possui uma população maior que o Centro-Oeste, mas praticamente a mesma população atendida por esgoto. Já o Norte também possui uma população total maior, mas um atendimento muito menor, enquanto o Sudeste se destaca em termos de atendimento e população total.

4.3. Análise de desempenho para os modelos por macrorregião (locais)

Considerando os atributos médios dos dois anos anteriores a cada ano t, obteve-se modelos locais, isto é, treinados separadamente para cada macrorregião. A metodologia aplicada foi exatamente a já descrita neste trabalho, onde a estratificação dos dados em conjuntos de treino, teste e OOT foi realizada dentro de cada macrorregião. Os resultados das avaliações para estes modelos são sumarizados na Tabela 5. De modo geral, observa-se que, diferentemente do modelo global, o desempenho varia menos entre as

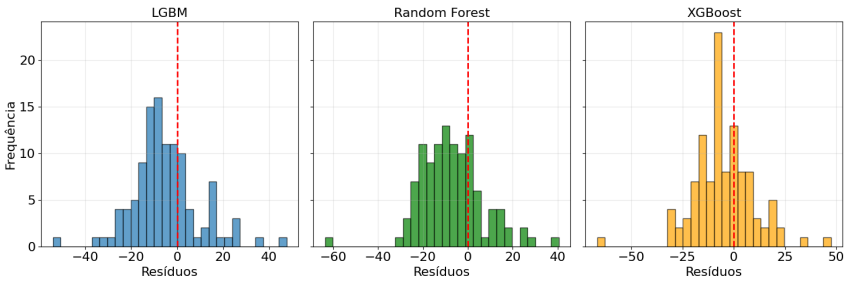
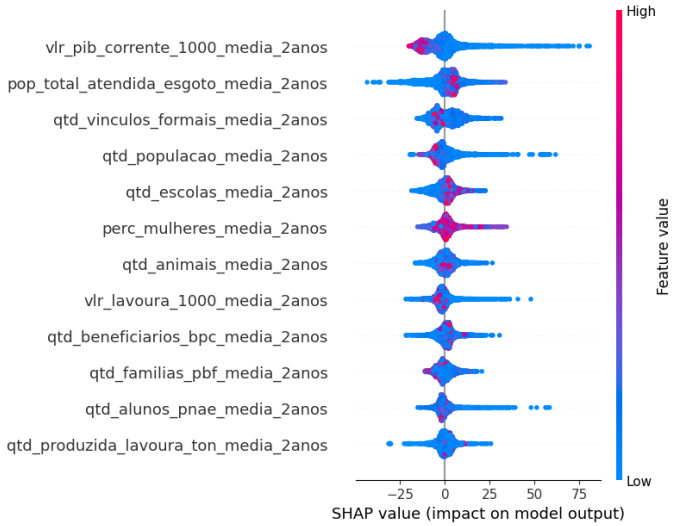


Figura 4. Histograma de resíduos no conjunto OOT para Sul

Figura 5. Análise SHAP para o modelo global (Brasil) baseado no *LGBM*.



macrorregiões, tantos em termos de R^2 quanto de MAPE. A macrorregião que apresenta a maior discrepância em relação às restantes é o Centro-Oeste, que possui métricas piores. Ressalta-se, porém, que, ao se treinar modelos específicos para esta macrorregião, seu desempenho subiu consideravelmente em relação ao seu desempenho no modelo global. No modelo global a melhor métrica MAPE para o Centro-Oeste foi de 0,54 (*LGBM*), já no modelo local a melhor métrica MAPE foi de 0,35 no *RF* e *XGBoost*.

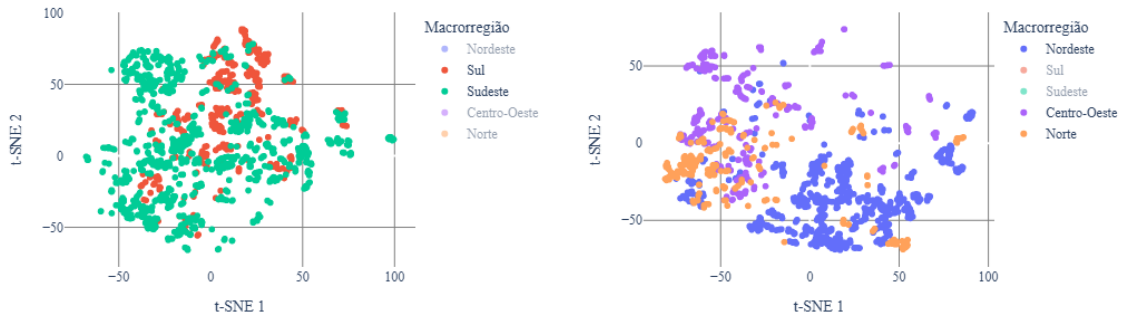
Em termos comparativos de métrica entre os modelos locais e globais, vê-se que o Centro-Oeste melhora, já o Sul piora, considerando-se o MAPE. No modelo global, o Sul tem sua melhor métrica no modelo global no valor de 0,25, tanto para LGBM quanto para *XGBoost*, já no modelo local o melhor valor é de 0,32 para *RF* e *XGBoost*. Uma hipótese a ser levantada é que, as macrorregiões Sul e Sudeste são, em certa medida, similares, o que pode colaborar para o desempenho do Sul no modelo global, tendo em vista que o Sudeste possui um quantitativo maior de regiões do SUS. Já as macrorregiões Norte, Nordeste e Sudeste possuem uma leve melhora de métrica.

A similaridade entrega as macrorregiões Sul e Sudeste são mostradas na Figura 6a, que apresenta as projeções obtidas por meio do t-SNE para os mesmos subconjuntos regionais. O t-SNE, por ser uma técnica não linear, evidencia agrupamentos locais definidos. Para as macrorregiões Norte, Nordeste e Centro-Oeste, apresentadas na Figura 6b, observa-se a formação de pequenos agrupamentos dispersos, com maior heterogeneidade

Tabela 5. Métricas dos modelos locais.

Modelo	*LGBM*		*RF*		*XGBoost*	
Região/Métrica	**R^2**	**MAPE**	**R^2**	**MAPE**	**R^2**	**MAPE**
Centro-oeste	0,28	0,43	0,26	0,35	0,21	0,35
Nordeste	0,53	0,31	0,55	0,34	0,57	0,31
Norte	0,53	0,31	0,55	0,34	0,57	0,31
Sudeste	0,54	0,31	0,55	0,34	0,56	0,31
Sul	0,52	0,33	0,57	0,32	0,43	0,32

para a região Nordeste. No caso das macrorregiões Sul e Sudeste, o t-SNE evidencia agrupamentos mais compactos e bem definidos, indicando maior similaridade entre as macrorregiões do SUS pertencentes a essas macrorregiões.



(a) Sul e Sudeste **(b) Norte, Nordeste e Centro-Oeste**

Figura 6. Visualização t-SNE das regiões de saúde brasileiras.

5. Conclusão

Os experimentos realizados mostraram que é possível obter capacidade preditiva para a taxa de internação por desnutrição a partir dos indicadores analisados. Os modelos são válidos, para além da predição da própria taxa, para entender quais os atributos que influenciam a taxa, uma vez que para se diminuir a desnutrição deve-se tratar a sua origem e não a desnutrição. Os valores de coeficiente de determinação obtidos, para o conjunto OOT, ficaram próximos a 0,65 e erros percentuais médios entre 35% e 42%. A análise por macrorregiões evidenciou, entretanto, diferenças expressivas no desempenho dos modelos, sugerindo que a relação entre determinantes socioeconômicos e internações por desnutrição não se manifesta de maneira homogênea no território nacional.

Entre as limitações do estudo, destacam-se a dependência da qualidade e da periodicidade das bases públicas utilizadas e possíveis subnotificações nos registros de internação. Como perspectivas futuras, sugere-se a avaliação da estabilidade do modelo para diferentes períodos de OOT, visando investigar efeitos de *concept drift*, a incorporação de novas fontes de dados, como indicadores de insegurança alimentar em nível domiciliar, dados de atenção primária e variáveis ambientais, além da investigação de arquiteturas baseadas em séries temporais profundas. Seria importante, também, analisar o acesso à rede pública de saúde, pois apesar da internação por desnutrição ser um *proxy* para a própria desnutrição, pode haver, ou não, um *gap* significativo entre o número de pessoas em estado de desnutrição e a internação por essa comorbidade.

Agradecimentos

Este estudo foi financiado, em parte, pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Referências

Almalki, A., Gokaraju, B., Mehta, N., and Doss, D. A. (2021). Geospatial and machine learning regression techniques for analyzing food access impact on health issues in sustainable communities. *ISPRS International Journal of Geo-Information*, 10(11).

Almeida, G., Brito Correia, F., Borges, A. R., and Bernardino, J. (2024). Hospital length-of-stay prediction using machine learning algorithms—a literature review. *Applied Sciences*, 14(22).

Awan, A. A. (2024). Introdução ao t-sne: Redução de dimensionalidade não linear e visualização de dados. Acessado em: 22 dez. 2025.

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.

Cao, G., Kornher, L., and Brandi, C. (2025). How robust are machine learning approaches for improving food security amid crises?-evidence from COVID-19 in Uganda. *World Development*, 196:107171.

da Silva, T. L. C., Furtado, L. S., Fernandes, G. S., de Macêdo, J. A. F., Cruz, L. A., Magalhães, R. P., and Gomes, L. G. A. (2025). Machine learning methods and models to predict food insecurity levels for families in Ceará, Brazil, based on employment, housing and other social indicators. *Journal of the Brazilian Computer Society*, 31(1):203–218.

FAO, IFAD, UNICEF, WFP, and WHO (2025). The state of food security and nutrition in the world 2025 – addressing high food price inflation for food security and nutrition. Technical report, Food and Agriculture Organization of the United Nations, Rome.

Jarray, N., Abbes, A. B., and Farah, I. R. (2023). Machine learning for food security: current status, challenges, and future perspectives. *Artificial Intelligence Review*, 56(Suppl 3):3853–3876.

Li, J. (2017). Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? then what? *PLOS ONE*, 12(8):1–16.

Simeon, Y. (2023). Alimentação escolar é a principal refeição para 56% dos estudantes do grande rio, revela pesquisa. Acessado em: 31 dez. 2025.

Valadares, A., Souza, M., and Esteves, M. (2023). Mapeamento da insegurança alimentar e nutricional com foco na desnutrição (mapa insan) a partir da análise do sistema nacional de vigilância alimentar e nutricional (sisvan). Technical report, Ministério do Desenvolvimento e da Assistência Social, Família e Combate à Fome Secretaria Extraordinária de Combate à Pobreza e à Fome.

Xiong, R., Peng, H., Chen, X., and Shuai, C. (2024). Machine learning-enhanced evaluation of food security across 169 economies. *Environment, Development and Sustainability*, 26(10):26971–27000.