UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ESPECIALIZAÇÃO EM ENGENHARIA DE SOFTWARE PARA APLICAÇÕES DE CIÊNCIA DE DADOS

VANESSA PAULA RESMINI

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

Monografia de Conclusão de Curso apresentada como requisito parcial para a obtenção do grau de Especialista em Engenharia de Software para Aplicações de Ciência de Dados

Orientadora: Prof[a]. Dr[a]. Mariana Recamonde Mendoza

Porto Alegre
2026

CIP — CATALOGAÇÃO NA PUBLICAÇÃO

Resmini, Vanessa Paula

Predição da taxa de internação por desnutrição no Sistema Único de Saúde utilizando aprendizado de máquina / Vanessa Paula Resmini. – Porto Alegre: PPGC da UFRGS, 2026.

66 f.: il.

Monografia (especialização) – Universidade Federal do Rio Grande do Sul. Curso de Especialização em Engenharia de Software para Aplicações de Ciência de Dados, Porto Alegre, BR–RS, 2026. Orientadora: Mariana Recamonde Mendoza.

1. Desnutrição. 2. Sistema Único de Saúde. 3. Aprendizado de máquina. 4. Modelos preditivos. I. Mendoza, Mariana Recamonde. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitora: Prof[a]. Márcia Cristina Bernardes Barbosa
Vice-Reitor: Prof. Pedro de Almeida Costa
Pró-Reitora Adjunta de Pós-Graduação: Prof[a]. Claudia Wasserman
Diretor do Instituto de Informática: Prof. Luciano Paschoal Gaspary
Coordenadores da Especialização: Prof[a]. Dra. Karin Becker e Prof. Dr. Marcelo Pimenta
Bibliotecário-chefe do Instituto de Informática: Alexsander Borges Ribeiro

RESUMO

O Brasil ainda luta para prover à população acesso a uma alimentação nutritiva e balanceada, fato reforçado pela reinclusão ao Mapa da Fome da Organização das Nações Unidas em 2019. Ainda que, em 2025, o Brasil tenha saído do Mapa da Fome, há trinta e cinco milhões de pessoas com dificuldade para se alimentar, demonstrando a necessidade em manter e estruturar as políticas públicas de redução de fome e pobreza. O Sistema Único de Saúde (SUS) possui um papel importante neste contexto, tanto na prevenção quanto na recuperação da saúde de pessoas afetadas pela fome. Neste trabalho, foram desenvolvidos e avaliados modelos de aprendizado de máquina com o objetivo de predizer a taxa anual de internação hospitalar por comorbidade de desnutrição no Sistema Único de Saúde, segmentada pelas 439 regiões de saúde do país, com base no local de residência dos pacientes. Para a construção dos modelos, foram utilizados dados históricos de internações do SUS combinados com variáveis socioeconômicas, demográficas, educacionais e de infraestrutura provenientes de bases públicas nacionais. Foram treinados e comparados os algoritmos *Random Forest*, *XGBoost* e *Light Gradient Boosting Machine* com aplicação de técnicas de seleção de atributos, otimização de hiperparâmetros e validação temporal do tipo *out-of-time* (OOT). O desempenho dos modelos foi avaliado por meio das métricas coeficiente de determinação (R^2) e erro percentual absoluto médio (MAPE). No conjunto OOT, observou-se valores de R^2 de 0,66 para LGBM, 0,64 para *Random Forest* 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 significativas no desempenho dos modelos entre as macrorregiões brasileiras, indicando que fatores socioeconômicos e territoriais influenciam de maneira distinta a taxa de internação por desnutrição, assim como foi possível observar na interpretação dos algoritmos, utilizando-se SHAP.

Palavras-chave: Desnutrição. Sistema Único de Saúde. Aprendizado de máquina. Modelos preditivos.

Prediction of hospitalization rate for malnutrition in the Sistema Único de Saúde (SUS) using machine learning

ABSTRACT

Brazil still struggles to provide its population with access to nutritious and balanced food, a fact reinforced by its reinclusion in the United Nations Hunger Map in 2019. Even if Brazil is removed from the Hunger Map by 2025, thirty-five million people experience food insecurity, demonstrating the need to maintain and structure public policies to reduce hunger and poverty. The Sistema Único de Saúde (SUS) plays an important role in this context, both in the prevention and the recovery of the health of people affected by hunger. In this work, machine learning models were developed and evaluated with the aim of predicting the annual rate of hospitalizations due to malnutrition in the Sistema Único de Saúde, segmented by the 439 health regions of the country, based on the patients place of residence. For the construction of the models, historical data on SUS hospitalizations were used, combined with socioeconomic, demographic, educational, and infrastructure variables from national public databases. The algorithms Random Forest, XGBoost, and Light Gradient Boosting Machine were trained and compared using feature selection techniques, hyperparameter optimization, and out-of-time (OOT) validation. Model performance was evaluated using the coefficient of determination (R^2) and mean absolute percentage error (MAPE). In the OOT set, 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 revealed significant variations in the performance of the models among the Brazilian macro-regions, indicating that socioeconomic and territorial factors influence the rate of hospitalization due to malnutrition in distinct ways, as was observed in the interpretability of the algorithms, using SHAP.

Keywords: Malnutrition. Sistema Único de Saúde. Machine Learning. Predictive models. Brazil.

LISTA DE ABREVIATURAS E SIGLAS

BPC	Benefício de Prestação Continuada
CAGED	Cadastro Geral de Empregados e Desempregados
CNAE	Classificação Nacional de Atividades Econômicas
EBIA	Escala Brasileira de Insegurança Alimentar
IBGE	Instituto Brasileiro de Geografia e Estatística
LGBM	*Light Gradient Boosting Machine*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
OOT	*Out-of-time*
PBF	Programa Bolsa Família
PCA	*Principal Component Analysis*
PIB	Produto Interno Bruto
PNAE	Programa Nacional de Alimentação Escolar
Q-Q	*Quantile-quantile*
RAIS	Relação Anual de Informações Sociais
RFE	*Recursive Feature Elimination*
RMSE	*Root Mean Squared Error*
SAGICAD	Secretaria de Avaliação, Gestão da Informação e Cadastro Único
SFM	*Select From Model*
SFS	*Sequential Feature Selector*
SHAP	*Shapley Additive Explanations*
SNIS	Sistema Nacional de Informações sobre Saneamento
SUS	Sistema Único de Saúde
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
XGBoost	*eXtreme Gradient Boosting*

LISTA DE FIGURAS

Figura 3.1 Exemplo de validação utilizando OOT 17
Figura 3.2 Exemplo de aplicação de SHAP 18

Figura 5.1 Histórico da taxa de internação por desnutrição em t-1 27
Figura 5.2 Distribuição das regiões do SUS nas macrorregiões do Brasil 28
Figura 5.3 População total 29
Figura 5.4 Valor do Produto Interno Bruto 30
Figura 5.5 Valor médio do salário formal de admissão no mês de dezembro 30
Figura 5.6 Pessoas atendidas pelo Programa Bolsa Família 31

Figura 6.1 Seleção de atributos usando RFE com LGBM 38
Figura 6.2 Seleção de atributos usando RFE com *Random Forest* 38
Figura 6.3 Seleção de atributos usando RFE com *XGBoost* 39
Figura 6.4 Q-Q para *Random Forest* 41
Figura 6.5 Q-Q para *Random Forest* com otimização 41
Figura 6.6 *Boxplot* da taxa de internação por macrorregião (em t-1) 45
Figura 6.7 Taxa média de internação por macrorregião 46
Figura 6.8 Resíduos vs. valores atuais no conjunto OOT para Centro-Oeste 46
Figura 6.9 Histograma de resíduos no conjunto OOT para Centro-Oeste 47
Figura 6.10 Histograma de resíduos no conjunto OOT para Sudeste 47
Figura 6.11 Histograma de resíduos no conjunto OOT para Sul 48
Figura 6.12 SHAP para Brasil (LGBM) 48
Figura 6.13 População total distribuída por macrorregião 49
Figura 6.14 População total atendida por esgoto distribuída por macrorregião 50
Figura 6.15 SHAP para Sudeste (LGBM) 50
Figura 6.16 SHAP para Sul (LGBM) 51
Figura 6.17 SHAP para Norte (LGBM) 52
Figura 6.18 SHAP para Nordeste (LGBM) 53
Figura 6.19 SHAP para Centro-Oeste (LGBM) 53
Figura 6.20 PCA (exceto macrorregiões Sul e Sudeste) 56
Figura 6.21 PCA (macrorregiões Sul e Sudeste) 56
Figura 6.22 t-SNE (exceto macrorregiões Sul e Sudeste) 57
Figura 6.23 t-SNE (macrorregiões Sul e Sudeste) 58

LISTA DE TABELAS

Tabela 5.1 Atributos explicativos do atributo-alvo.....25
Tabela 5.2 Hiperparâmetros para LGBM, *Random Forest* e *XGBoost* 35

Tabela 6.1 Métricas do modelo *baseline* baseado em regressão linear.....37
Tabela 6.2 Atributos selecionados para cada modelo 39
Tabela 6.3 Métricas dos modelos após seleção de atributos 40
Tabela 6.4 Métricas dos modelos após seleção de atributos e busca por hiperparâmetros40
Tabela 6.5 Atributos selecionados para cada modelo após adição de variáveis.....42
Tabela 6.6 Métricas dos modelos após adição de atributos e busca por hiperparâmetros43
Tabela 6.7 Métricas dos modelos após categorização e busca por hiperparâmetros.....44
Tabela 6.8 Métricas da Tabela 6.6 abertas por macrorregião.....45
Tabela 6.9 Métricas dos modelos locais.....54
Tabela 6.10 Atributos que mais contribuem para PCA.....57

Tabela 8.1 Atributos selecionados para cada modelo para o Centro-Oeste 61
Tabela 8.2 Atributos selecionados para cada modelo para o Nordeste 62
Tabela 8.3 Atributos selecionados para cada modelo para o Norte 62
Tabela 8.4 Atributos selecionados para cada modelo para o Sudeste 63
Tabela 8.5 Atributos selecionados para cada modelo para o Sul 63

SUMÁRIO

1 INTRODUÇÃO 9
2 OBJETIVOS 11
2.1 Objetivo principal 11
2.2 Objetivos secundários 11
3 FUNDAMENTAÇÃO TEÓRICA 12
3.1 Aprendizado de máquina 12
3.2 Técnicas para seleção de atributos 13
3.3 Otimização de hiperparâmetros 14
3.4 Avaliação de modelos 15
3.4.1 Métricas de avaliação 15
3.4.2 Validação cruzada 16
3.4.3 Avaliação em séries temporais 16
3.5 Explicabilidade de modelos 17
3.6 Redução de dimensionalidade 18
4 REVISÃO DA LITERATURA 20
5 METODOLOGIA 23
5.1 Coleta e pré-processamento dos dados 23
5.1.1 Histórico da internação por desnutrição 27
5.1.2 Análise exploratória dos atributos 28
5.2 Montagem e treinamento dos modelos 30
5.2.1 Pré-processamento para o treinamento 31
5.2.2 Seleção de atributos 32
5.2.3 Otimização de hiperparâmetros 32
5.2.4 Categorização 33
5.3 Explicabilidade 33
5.4 Implementação prática 34
6 RESULTADOS E DISCUSSÃO 37
6.1 Modelo *baseline* 37
6.2 Seleção de atributos 37
6.3 Otimização bayesiana 40
6.4 Adição de atributos 41
6.5 Categorização de atributos contínuos 43
6.6 Desempenho dos modelos nas macrorregiões do país 44
6.7 Modelos locais para as macrorregiões do país 53
6.8 Redução de dimensionalidade 55
7 CONCLUSÃO 59
8 APÊNDICE 61
REFERÊNCIAS 64

1 INTRODUÇÃO

A desnutrição é um indicador de vulnerabilidade social e sanitária que reflete não apenas a insuficiência alimentar, mas também a precariedade das condições de vida, do acesso a serviços públicos e da eficácia das políticas sociais, conforme apontado por Valadares, Souza and Esteves (2023) no Mapeamento da Insegurança Alimentar e Nutricional com foco na Desnutrição. 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, conforme apontado por FAO et al. (2025), ainda há trinta e cinco milhões de pessoas com dificuldade para se alimentar.

A taxa de internação hospitalar pela morbidade de desnutrição no Sistema Único de Saúde (SUS) oferece um panorama relevante sobre os impactos da insegurança alimentar e das desigualdades territoriais. A compreensão dos fatores que influenciam esse indicador é essencial para o planejamento de ações intersetoriais voltadas à promoção da saúde e à redução das desigualdades, uma vez que a insegurança alimentar e a desnutrição não se apresentam de maneira igualitária no Brasil, mas, sim, associadas às desigualdades regionais e econômicas, assim como apontado por Valadares, Souza and Esteves (2023) e Wissenbach, Bicev and Vanetti (2024).

Com o avanço das tecnologias de análise de dados, especialmente os modelos de aprendizado de máquina, torna-se possível explorar dados públicos e identificar padrões complexos entre variáveis socioeconômicas, educacionais, demográficas e de infraestrutura. Assim, este trabalho propõe o treinamento de modelos para predizer a taxa de internação hospitalar por comorbidade de desnutrição, além de compreender os seus determinantes, contribuindo para o aprimoramento das políticas públicas de saúde e de segurança alimentar. Adicionalmente, objetiva-se realizar um estudo comparativo entre as diferentes macrorregiões do Brasil em termos de poder preditivo dos modelos e de variáveis relevantes na tarefa de predição.

O presente trabalho está organizado em sete capítulos. No Capítulo 2, são apresentados o objetivo principal e os objetivos secundários da pesquisa. O Capítulo 3 reúne a fundamentação teórica necessária, abordando conceitos de aprendizado de máquina, técnicas de validação, seleção de atributos, otimização de hiperparâmetros, métricas de avaliação e métodos de explicabilidade. No Capítulo 4, é realizada a revisão da literatura, destacando estudos sobre desnutrição, insegurança alimentar e modelagem preditiva em

saúde pública. O Capítulo 5 descreve detalhadamente a metodologia adotada, incluindo a coleta e o pré-processamento dos dados, a construção dos modelos, os procedimentos de seleção de atributos, de otimização e de explicabilidade. No Capítulo 6, são apresentados e discutidos os resultados obtidos, com ênfase no desempenho dos modelos e na interpretação dos fatores associados à taxa de internação por desnutrição. Por fim, no Capítulo 7, são apresentadas as considerações finais.

2 OBJETIVOS

2.1 Objetivo principal

Analisar e prever a taxa anual de internação hospitalar por morbidade associada à desnutrição no sistema público de saúde, segmentada por regiões do Sistema Único de Saúde (SUS), com base em variáveis socioeconômicas, educacionais, demográficas e de infraestrutura, utilizando técnicas de aprendizado de máquina.

De acordo com o Ministério da Saúde, as regiões de saúde são formadas por municípios fronteiriços que compartilham identidades culturais, econômicas e sociais, redes de comunicação e infraestrutura de transportes. A finalidade dessas regiões é integrar a organização, o planejamento e a execução de ações e serviços de saúde.

2.2 Objetivos secundários

Os objetivos secundários deste trabalho consistem em aprofundar a compreensão dos fatores associados à taxa de internação por desnutrição no sistema público de saúde. Para isso, foram definidas etapas analíticas que envolvem a comparação de modelos preditivos, a identificação de variáveis relevantes por meio de técnicas de seleção de atributos e a investigação da influência da segmentação territorial. Esses objetivos específicos são detalhados a seguir:

- Comparar diferentes modelos preditivos aplicados à modelagem das taxas de internação por desnutrição no sistema público de saúde, avaliando simultaneamente o desempenho preditivo e a capacidade explicativa frente à complexidade dos dados;
- Identificar variáveis relevantes associadas à taxa de internação por desnutrição por meio da aplicação de técnicas de seleção de atributos, visando lidar com a alta dimensionalidade do conjunto de dados e compreender os fatores mais fortemente relacionados ao fenômeno estudado;
- Analisar o impacto da segmentação territorial na variabilidade das taxas de internação, considerando tanto as regiões de saúde quanto as macrorregiões brasileiras;
- Investigar o papel do espaço geográfico como determinante social da saúde, examinando de que forma desigualdades regionais refletem nos padrões de internação hospitalar por desnutrição.

3 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão abordados os principais conceitos relacionados ao desenvolvimento deste trabalho, incluindo uma breve revisão sobre aprendizado de máquina supervisionado, alguns algoritmos selecionados, técnicas para seleção de variáveis e de modelos, além de abordagens para explicabilidade.

3.1 Aprendizado de máquina

De acordo com Nichols, Chan and Baker (2018), aprendizado de máquina é um termo que se refere a uma gama de algoritmos que são capazes de predizer baseado em um conjunto de dados. Há uma variedade de algoritmos de aprendizado de máquina, normalmente referidos sob o termo "modelo", que são amplamente utilizados atualmente. Dentro do escopo destes modelos, estão os de classificação e os de regressão, sendo estes os abordados neste trabalho.

Os modelos de regressão são um subgrupo dos algoritmos de aprendizado supervisionado, em que o treinamento é realizado com base em um conjunto de dados rotulados. Nesses modelos, a variável dependente y é resultado dos valores das variáveis independentes x. As técnicas de regressão têm, majoritariamente, dois usos: o de predição da variável dependente; e o de estabelecer as relações causais entre as variáveis independentes e as dependentes (MAULUD; ABDULAZEEZ, 2020). Ambos os usos serão explorados neste trabalho.

Dentre os diversos modelos de regressão existentes, foram abordados árvore de decisão, *eXtreme Gradient Boosting* (*XGBoost*), *Light Gradient Boosting Machine* (LGBM) e *Random Forest*, os quais têm ganhado popularidade devido à sua robustez e ao alto poder de predição (KUMAR; SRIVASTAVA; PRAKASH, 2024). Os quatro modelos são algoritmos supervisionados que podem ser utilizados tanto para classificação como para regressão. Enquanto a saída de uma classificação é a classe correspondente, na regressão, é um valor contínuo ou inteiro. Os algoritmos são descritos a seguir:

- Árvore de decisão: é um algoritmo que se utiliza de uma estrutura semelhante a de uma árvore. Ele divide os dados baseado em características, partindo de uma pergunta raiz e ramificando-se a partir dela. Cada nó pergunta sobre um atributo, dividindo os dados até chegar aos nós com as previsões finais (KHARKAR, 2020).

O uso de árvores de decisão para regressão são menos populares do que seu uso para classificação, no entanto possuem métricas competitivas e usualmente são capazes de atender problemas da vida real (CZAJKOWSKI; KRETOWSKI, 2016). Entre as desvantagens, estão o *overfitting*, que ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, levando à superestimação de seu desempenho e, consequentemente, à sua ineficácia em dados de teste e à instabilidade em algumas situações, porém são de fácil interpretação e muito úteis em exploração de dados;

- *Random Forest*: é um modelo de aprendizado que funciona criando um conjunto de árvores de decisão independentes a partir de diferentes subconjuntos do conjunto de dados original. O modelo final é uma combinação das previsões de cada árvore (KUMAR; SRIVASTAVA; PRAKASH, 2024);
- *Extreme Gradient Boosting*: é uma técnica que combina predições de outros modelos, frequentemente árvores de decisão, para construir um modelo mais robusto (KUMAR; SRIVASTAVA; PRAKASH, 2024). Nesse caso, cada nova árvore é treinada para corrigir os erros da árvore anterior, em um processo chamado de *boosting*, com o objetivo de diminuir o gradiente descendente. Esta técnica vem sendo utilizada por muitos cientistas de dados para atingir excelentes resultados em diferentes problemas da vida real, como detecção de fraude, detecção de anomalias, etc (CHEN; GUESTRIN, 2016);
- *Light Gradient Boosting Machine*: LGBM é um algoritmo de aprendizado desenvolvido pela Microsoft, que, assim como o *XGBoost*, busca diminuir o gradiente descendente utilizando-se da combinação de múltiplas árvores de decisão. Ele foi projetado para lidar com grande volume de dados, performando bem em termos de velocidade e uso de memória (KUMAR; SRIVASTAVA; PRAKASH, 2024).

3.2 Técnicas para seleção de atributos

Seleção de atributos é uma estratégia de pré-processamento que já se provou efetiva para diversos problemas de mineração de dados e de aprendizado de máquina, especialmente em conjuntos de dados de alta dimensionalidade (LI et al., 2017). Selecionar os atributos mais adequados melhora a performance dos modelos, assim como diminui a probabilidade de *overfitting*, reduz a quantidade de dados, o que torna o treino mais rápido, além de tornar o modelo mais fácil de interpretar.

Dentro das técnicas de seleção supervisionadas, há três categorias: a de filtragem, que considera apenas os dados e não um modelo específico, classificando cada atributo em relação ao atributo-alvo conforme alguma métrica estatística, dentro desta categoria se enquadra o método *information value*; o *wrapper*, que utiliza um subconjunto dos dados para treinar algum modelo e, baseado na performance, adiciona ou remove atributos, como o *recursive feature elimination* (RFE); e, por fim, os métodos *embedded*, os quais combinam a seleção ao próprio processo de treinamento. Quando a performance do treinamento diminui, são utilizados mecanismos para melhorar as próximas iterações, como regularização (BELCIC; STRYKER, 2025).

Dentre os métodos mais utilizados para seleção de atributos, o RFE funciona especialmente bem em amostras pequenas, tendendo a remover atributos redundantes e fracos (CHEN; JEONG, 2007). Nesta técnica, um modelo é treinado com todos os atributos, os quais vão sendo eliminados em cada iteração de acordo com a sua importância relativa, onde o menos importante é eliminado até que se tenha o número de atributos desejado ou se tenha a melhor métrica de performance. Este método pode ser combinado com a validação cruzada, descrita na Seção 3.4, na qual a métrica é calculada como a média entre todas as divisões, e selecionam-se os atributos que otimizam esse valor médio.

3.3 Otimização de hiperparâmetros

Otimização de hiperparâmetros é a prática de selecionar os melhores hiperparâmetros para cada modelo para cada conjunto de dados através de iterações. Hiperparâmetros são as configurações de cada modelo que podem ser modificadas a fim de controlar o processo de treino e que não podem ser inferidas diretamente dos dados, assim devem ser escolhidas antes do treino. É um processo crítico, o qual impacta diretamente a performance dos modelos.

Frequentemente, os hiperparâmetros são escolhidos manualmente ou através de uma combinação de pesquisa aleatória ou testando todas as combinações possíveis, o que requer muito tempo e processamento (WISTUBA; SCHILLING; SCHMIDT-THIEME, 2015). No entanto, há soluções que buscam otimizar a definição dos hiperparâmetros, como a otimização bayesiana.

Otimização bayesiana é uma classe de métodos de otimização baseado em aprendizado de máquina que tem o objetivo de maximizar uma função, onde a escolha dos hiperparâmetros de cada iteração é baseada no resultado da iteração anterior, sendo a nova

escolha provavelmente melhor do que a anterior. Essa técnica específica pertence a um grupo chamado de *surrogate methods*, nos quais é mantido um *surrogate* que representa a função objetivo. Esta técnica é muito mais rápida e eficiente do que outros métodos, pois não testa combinações não adequadas (FRAZIER, 2018).

Especificamente, a otimização bayesiana utiliza *surrogates* desenvolvidos utilizando estatística bayesiana, na qual a probabilidade é uma medida de incerteza e não é necessário atribuir uma distribuição de probabilidade aos parâmetros.

3.4 Avaliação de modelos

3.4.1 Métricas de avaliação

Modelos, sejam de classificação ou de regressão, não são perfeitamente precisos e não devem ser, para evitar o problema de *overfitting*. Assim, existem diversas métricas para avaliação e comparação de modelos de regressão, sendo as mais utilizadas *root mean squared error* (RMSE), *mean absolute error* (MAE) e o coeficiente de determinação (R^2), embora muitas outras existam e sejam utilizadas nas mais diversas áreas. No geral, o objetivo das métricas de performance é medir o desvio das predições em relação às observações reais, logo, quanto maior a diferença entre o real e o predito, mais distante o modelo está de ser uma representação acurada da realidade (PLEVRIS et al., 2022).

Uma das métricas escolhidas para este trabalho foi o R^2, seguindo outros trabalhos de cunho semelhante, como Almalki et al. (2021), Almalki et al. (2022) e Almeida et al. (2024). O coeficiente de determinação é uma unidade de medida adimensional que indica a proporção de variabilidade da variável dependente y que é explicada pelas variáveis independentes x. R^2 é dado por

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \bar{y_i})^2} \tag{3.1}$$

onde $\widehat{y_i}$ é o valor predito, y_i é o valor real e $\bar{y}$ é o valor médio da variável dependente y.

O R^2 é bastante popular, porém recebe críticas de alguns autores, como Li (2017), pois esta métrica olha apenas o comportamento das variáveis em relação uma a outra, sem considerar seus valores. 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, como o *mean absolute percentage error* (MAPE), que calcula um erro percentual, tornando mais fácil o comparativo entre diferentes modelos. A fórmula é mostrada na equação 3.2

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \cdot 100 \tag{3.2}$$

em que n é o número de pontos preditos, y_i e $\widehat{y_i}$ são, respectivamente, os valores reais e previstos para a instância i.

3.4.2 Validação cruzada

Validação cruzada é uma das técnicas mais utilizadas para avaliação e seleção de modelos, sendo também utilizada para selecionar hiperparâmetros (BERRAR, 2018). É uma maneira de prever o poder de generalização dos modelos e de prevenir *overfitting*. Em linhas gerais, consiste em treinar os modelos em diversos subconjuntos de dados, avaliando os resultados nestes diferentes conjuntos, usualmente através da média do desempenho entre esses subconjuntos. Existem diversas variações desta técnica de acordo com a forma com a qual os subconjuntos de dados são criados.

3.4.3 Avaliação em séries temporais

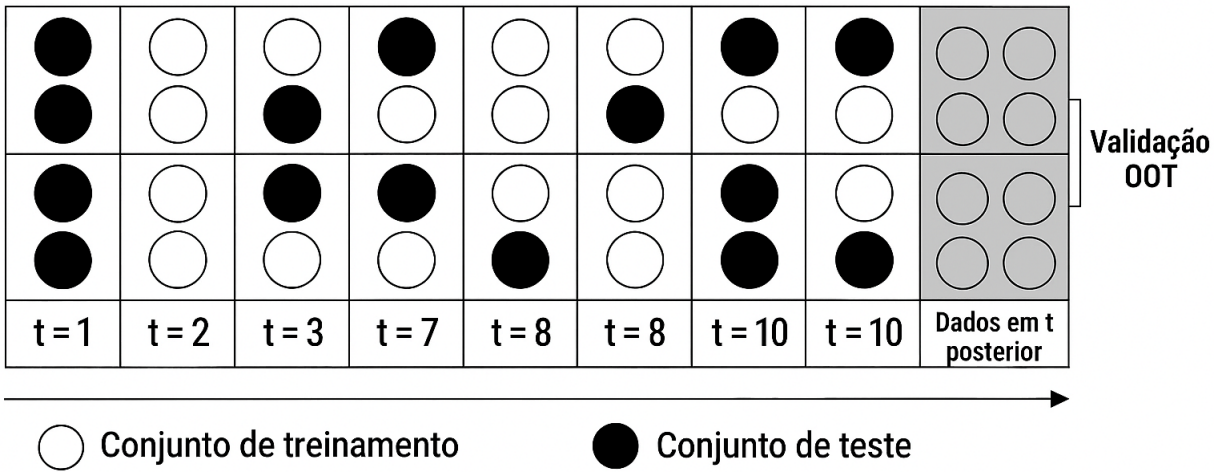
Uma parte fundamental da avaliação de um projeto de aprendizado de máquina é avaliar a performance em dados que não foram utilizados para treino. Embora, conforme mencionado, a validação cruzada seja utilizada para evitar *overfitting*, no caso de séries temporais, este método não preserva a ordem temporal das observações, o que pode levar a uma superestimação da performance (CERQUEIRA; TORGO; MOZETIč, 2020). De acordo com Hibbeln et al. (2025), utilizar validação *out-of-time* (OOT) resulta em uma melhor seleção de modelos, uma vez que preserva a ordem temporal das observações.

A validação OOT se baseia em separar um ou mais períodos de tempo anteriores à predição em produção apenas para validação, sem utilizar esses períodos t para treino. Isso impede que se utilize informação futura para prever o passado, por exemplo, se utilizando de valores em t_2 e t_3 para prever em t_1.

A Figura 3.1 ilustra um exemplo de utilização de validação OOT, onde há um conjunto de *t's* utilizados para treino e teste dentro do período selecionado para treinamento

do modelo e há um conjunto em t_{11} utilizado para validação OOT, em um período de tempo não utilizado em treino e que reproduz, de maneira mais adequada, o comportamento em um t no futuro.

Figura 3.1: Exemplo de validação utilizando OOT



Fonte: Adaptado de Hibbeln et al. (2025)

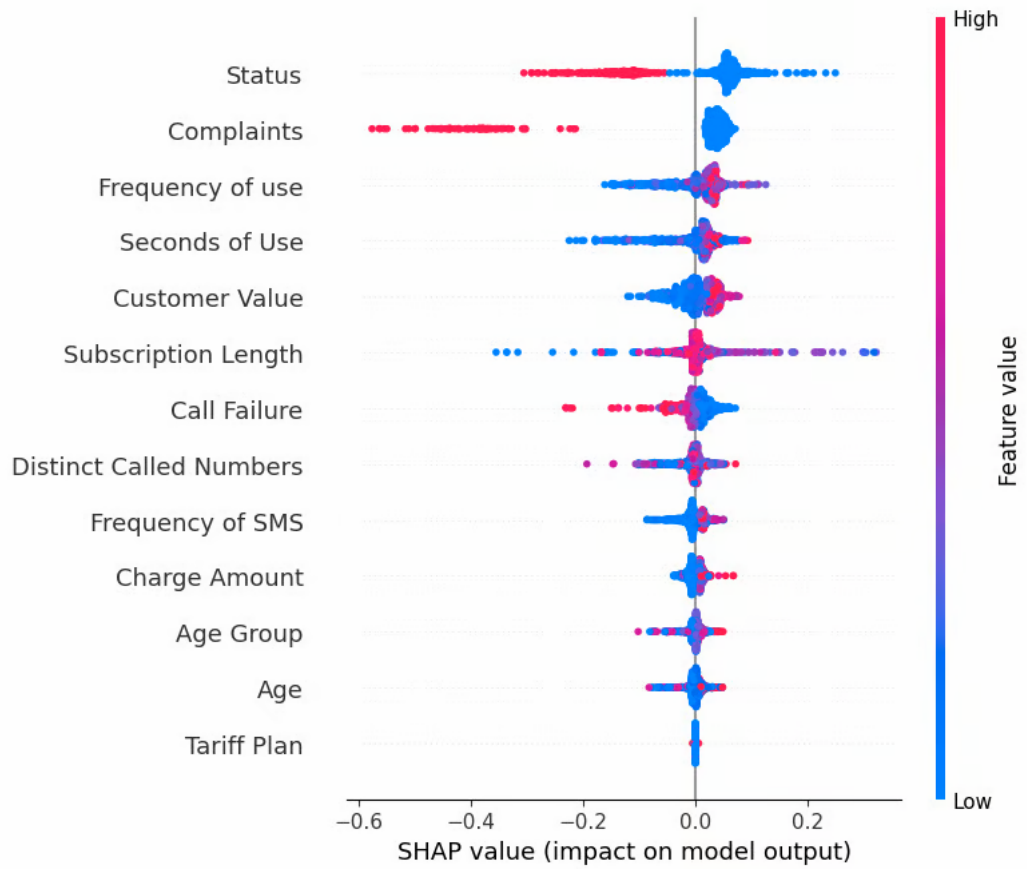
3.5 Explicabilidade de modelos

Explicabilidade de modelos trata sobre explicar e/ou justificar as saídas de modelos, frequentemente após as predições. Dentre as técnicas que podem ser utilizadas para auxiliar nisto, um dos métodos mais populares é o *shapley additive explanations* (SHAP) (SALIH et al., 2024). Este método é baseado na teoria dos jogos e considera cada atributo como um jogador e a saída do modelo como pagamento; ele atribui um valor (*shapley value*) a cada atributo baseado em quanto ele contribui para a predição em diferentes cenários. O SHAP pode ser aplicado tanto localmente quanto globalmente, ou seja, tem a habilidade de explicar o papel dos atributos para todas as instâncias ou para uma em específico. No SHAP, a ordem dos atributos representa a ordem de importância na saída do modelo, mas o valor não representa o peso dos atributos na predição final.

Um exemplo de uma avaliação SHAP para um modelo global, o qual classifica a evasão dos clientes de uma empresa de telecomunicações, pode ser visualizada na Figura 3.2 para a classe 0 (não evade). Podemos observar que, por exemplo, no atributo *complaints* (reclamações), quanto mais alto o valor, mais negativo o valor de SHAP, ou seja, quanto mais alta a quantidade de reclamações mais impacta negativamente a classe 0 e impacta positivamente a classe 1 (evade). Em linhas gerais, isso significa que quanto

maior o número de reclamações maior a probabilidade de evasão.

Figura 3.2: Exemplo de aplicação de SHAP



Fonte: Awan (2023)

3.6 Redução de dimensionalidade

A "maldição da dimensionalidade" é uma expressão utilizada para se referir ao rápido aumento e expansão de dimensões. Conjuntos de dados com esta característica podem conter informações de muita utilidade, porém a alta dimensionalidade, por si só, torna os dados complexos de serem analisados e estudados. Além disso, quando a dimensão dos dados aumenta, a performance dos algoritmos tende a melhorar, porém, quando aumenta ainda mais, a performance tende a piorar (JIA et al., 2022). Uma maneira de se analisar este tipo de dado é reduzir a sua dimensionalidade, mapeando os dados de um espaço com muitas dimensões para um espaço com menos dimensões. Dentro deste contexto, duas técnicas foram utilizadas, o *Principal Component Analysis* (PCA) e o *t-Distributed Stochastic Neighbor Embedding* (t-SNE).

O PCA é um método linear que transforma o conjunto original dos atributos em

um novo conjunto, menor do que o anterior, de variáveis não correlacionadas. Este conjunto é denominado de componentes principais; trata-se de combinações lineares das variáveis originais que têm a variação máxima em comparação com outras combinações lineares. Esses componentes buscam capturar o máximo possível de informações dos dados originais, porém eles só capturam comportamentos lineares, então conjuntos com relações mais complexas podem apresentar resultados menos satisfatórios (IBM, 2024).

Já o t-SNE é uma técnica não linear de redução de dimensionalidade, sendo útil para identificar agrupamentos que não são capturados por métodos lineares. Nesta técnica, são preservadas as relações entre os pontos de dados em um espaço de dimensão inferior, se preocupando em preservar pequenas distâncias de pares, diferente do PCA, que preserva grandes distâncias, com o objetivo de maximizar a variação (AWAN, 2024).

4 REVISÃO DA LITERATURA

A desnutrição é uma condição clínica e social que reflete a insuficiência alimentar crônica, frequentemente associada à pobreza, à desigualdade e à fragilidade das políticas públicas. No Brasil, a taxa de internação por desnutrição no SUS, segmentada pelas regiões de saúde do próprio SUS, pode representar um indicador epidemiológico sensível à insegurança alimentar e à exclusão social. A análise desse indicador, por meio de técnicas estatísticas e computacionais, pode tornar possível compreender certos determinantes sociais da saúde e orientar intervenções mais eficazes.

A insegurança alimentar é definida como a limitação no acesso regular e permanente a alimentos em quantidade e em qualidade adequadas. Estudos como o de Silva et al. (2025) demonstram que a insegurança alimentar pode ser prevista com elevada acurácia por meio de variáveis socioeconômicas, mesmo na ausência de dados nutricionais diretos. A pesquisa realizada com famílias do programa Cartão Mais Infância Ceará 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.

O estudo utilizou dados de 33.595 famílias em situação de vulnerabilidade. A partir de informações reportadas pelas famílias em uma pesquisa, foi construído um conjunto com vinte e nove atributos para o treinamento de diversos modelos para classificação. Os atributos foram organizados em cinco grandes dimensões analíticas:

- Insegurança alimentar: questões da Escala Brasileira de Insegurança Alimentar (EBIA), usadas para construir o rótulo do atributo-alvo;
- Acesso a alimentos: impacto da COVID-19 no acesso à comida e existência de produção doméstica (hortas ou criação de animais);
- Emprego e renda: recebimento de benefícios sociais; situação de emprego dos membros da família; número de pessoas ocupadas e renda familiar mensal;
- Habitação: tipo de posse do domicílio; material das paredes; acesso à água; presença de banheiro; coleta de lixo; iluminação; e localização em área de risco;
- Serviços públicos: disponibilidade de espaços de lazer, de atividades culturais e de participação em centros comunitários.

Houve a criação de dois cenários:

- Classificação multiclasse: quatro níveis de insegurança alimentar (severa, mode-

rada, leve e ausente);

- Classificação binária: insegurança severa e não severa.

Foram testadas diversas técnicas para seleção de atributos, dentre elas *Select From Model* (SFM), RFE, *Sequential Feature Selector* (SFS), Boruta e *Select K Best*. Cada método foi utilizado para avaliar a relevância dos atributos na tarefa de classificar os níveis de insegurança alimentar. A busca por hiperparâmetros foi realizada por um método de *grid search*. Ao final, considerando um conjunto de dados balanceado no cenário multiclasse, o modelo com melhor desempenho foi um *stacked ensemble* com seleção de atributos RFE e métrica F1 de 0,56. Já no cenário binário, sem balanceamento de classes, o melhor modelo foi um *distributed random forest*, com métrica F1 igual a 0,80. Ao se analisar a importância dos atributos para os melhores modelos, tanto na classificação multiclasse quanto na binária, o grupo de atributos "emprego e renda" mostrou-se o mais relevante, acumulando um pouco mais de cinquenta por cento do total de *feature importance*.

O uso de técnicas de aprendizado de máquina na saúde pública tem se consolidado como uma abordagem eficaz para a análise de grandes volumes de dados e para a previsão de desfechos clínicos. Conforme K and Sarathambekai (2021), algoritmos como *Random Forest, Lasso Regression, XGBoost* e LGBM têm demonstrado elevado desempenho na previsão de eventos de saúde. *Random forest*, por exemplo, foi utilizado para classificar câncer de mama com acurácia de 96,42%, sendo comumente usado para este problema, além de predição para doença cardiovascular e diabetes, pois, usualmente, consegue lidar bem com conjuntos de dados desbalanceados.

O uso de aprendizado de máquina na investigação e na previsão de desfechos relacionados à desnutrição tem crescido de forma expressiva nos últimos anos, especialmente em aplicações de saúde pública alicerçadas em grandes bases de dados populacionais. Uma meta-análise recente conduzida por Rao et al. (2025) sintetizou evidências de onze estudos que aplicaram algoritmos de aprendizado de máquina a dados das *Demographic and Health Surveys* para predizer diferentes formas de desnutrição infantil. Os autores observaram desempenho preditivo moderado a alto, com acurácia média de dez estudos sobre atraso no crescimento derivado de má nutrição, de 68,9%, já três estudos sobre emaciação, síndrome caracterizada pela perda não intencional de mais de 10% da massa basal e atrofia muscular, tiveram a acurácia média de 84,4%. O algoritmo regressão logística foi frequentemente utilizada para classificação binária, como para determinar se atraso de crescimento, emaciação ou baixo peso estavam presentes. Por fim, a meta-análise conclui que modelos de aprendizado de máquina e algoritmos baseados em árvores são mais

efetivos do que modelos lineares e outas técnicas, possivelmente por capturarem melhor as relações complexas entre determinantes socioeconômicos e desnutrição. Além disso, a revisão destaca o uso recorrente de técnicas de seleção de atributos e de métodos de explicabilidade, como SHAP, para identificar os principais fatores associados à desnutrição infantil.

A modelagem preditiva permite identificar padrões ocultos, explorar interações não lineares entre variáveis e realizar previsões com maior precisão do que os métodos estatísticos tradicionais. No contexto da desnutrição, essa abordagem é particularmente relevante, dada a complexidade dos fatores envolvidos — que incluem indicadores econômicos, educacionais, demográficos e de infraestrutura.

A taxa de internação por desnutrição, portanto, não deve ser interpretada apenas como um desfecho clínico, mas como um reflexo das condições estruturais que afetam o bem-estar da população. De acordo com Wissenbach, Bicev and Vanetti (2024), as regiões Norte e Nordeste, possuem, percentualmente, mais pessoas em situação de insegurança alimentar do que a média brasileira. É pontuado que, em 2023, a extrema pobreza atingia 1,7% dos brasileiros, mas afetava 2,7% dos nordestinos, já o desemprego no Nordeste era de 11%, enquanto no Brasil era 7,8%. O Norte e Nordeste ainda sofrem com a herança da concentração econômica no século XX, o que, ainda de acordo com Wissenbach, Bicev and Vanetti (2024), tem sido atenuada mais recentemente com as melhoras nos indicadores de educação e de saneamento básico.

5 METODOLOGIA

Neste capítulo, são apresentadas as etapas metodológicas adotadas no desenvolvimento do estudo, incluindo a coleta e a preparação dos dados, a construção e o treinamento dos modelos, as técnicas de seleção de atributos, os procedimentos de explicabilidade e as análises estatísticas realizadas.

5.1 Coleta e pré-processamento dos dados

O atributo-alvo escolhido para estudo é a taxa de internação hospitalar anual no sistema público de saúde pela morbidade de desnutrição, com base no local de residência dos pacientes, sendo esta uma informação disponível no DATASUS. A taxa é calculada por 100.000 pessoas e agregada nas regiões do SUS.

A escolha da taxa de internação por desnutrição como variável dependente é justificada por sua relevância epidemiológica e por sua capacidade de sintetizar múltiplos fatores de risco. Trata-se de um indicador que reflete não apenas o estado nutricional da população, mas também a eficácia das políticas públicas, a qualidade da infraestrutura sanitária e o acesso aos serviços de saúde.

Ao segmentar essa taxa por regiões de saúde do SUS, a análise incorpora uma dimensão territorial que reconhece o papel do espaço geográfico como determinante social da saúde. Essa abordagem permite identificar desigualdades e orientar intervenções mais precisas, conforme as necessidades locais. Conforme o Ministério da Saúde, a própria finalidade dessas regiões é integrar a organização, o planejamento e a execução de ações e serviços de saúde.

Para as escolhas dos atributos a serem estudados, baseou-se nas variáveis que Silva et al. (2025) sugere na predição de insegurança alimentar no estado do Ceará. De acordo com o artigo, o *dataset* deve representar as condições sociais e econômicas das regiões do SUS. Para este trabalho, o *dataset* foi montado com dados abertos provenientes de diferentes órgãos ou autarquias governamentais, abrangendo identificação geográfica, indicadores econômicos, demográficos, sociais, educacionais e de infraestrutura com informação temporal de 2012 a 2021, sendo 2021 o ano final do conjunto de dados, por ser o mais recente de algumas informações disponibilizadas.

A Tabela 5.1 abrange os atributos a serem analisados e suas origens. Todos os dados são fornecidos na granularidade de municipalidade, ou seja, em 2012 seriam 5.568

instâncias; por sua vez, estas instâncias são agrupadas nas 439 regiões através de soma, com exceção do percentual de mulheres, do percentual de adesão à política de saneamento e do salário médio de admissão, sob 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. Salário formal é aquele vindo de um vínculo celetista, quando o trabalhador possui registro na CLT, ou de um vínculo de servidor público estatutário, quando há registro na Relação Anual de Informações Sociais (RAIS).

Foram escolhidos atributos como a quantidade de alunos atendidos pelo Programa Nacional de Alimentação Escolar (PNAE), que fornece alimentos em escolas de ensino público do Brasil, além de programas de transferência de renda, como o Programa Bolsa Família (PBF) e o seu substituto temporário, o Auxílio Brasil, e o de assistência social, como o Benefício de Prestação Continuada (BPC), que tem como objetivo atender idosos e pessoas com deficiência. No aspecto econômico, também foram escolhidos atributos como o Produto Interno Bruto (PIB) e o salário de admissão dos vínculos formais, os quais são indicadores que buscam explicar o poder de compra da população. Já a quantidade de escolas e alunos abrangem dois aspectos, tanto o da renda, uma vez que quanto mais educada a população é maior é a renda média, como de acesso ao alimento, através da merenda escolar. Já outros atributos, como a da produção agrícola e animal e a quantidade de supermercados, são ligados à facilidade de acesso e de preço a um alimento saudável e nutritivo. Para a produção animal, foram escolhidos os bovinos, galináceos e suínos, que não sejam matrizes, mas animais de corte.

A quantidade de supermercados foi definida de acordo com a Classificação Nacional de Atividades Econômicas (CNAE) principal das empresas ativas durante o ano. Os CNAE's escolhidos foram:

- 471130: comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - hipermercados;
- 4711302: comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - supermercados;
- 4712100: comércio varejista de mercadorias em geral, com predominância de produtos alimentícios - minimercados, mercearias e armazéns.

O valor e quantidade de produção agrícola do ano foi extraída da Pesquisa Agrícola Municipal para lavouras permanentes e temporárias. Foram escolhidos os produtos de lavoura que são alimentos ou comestíveis, conforme citado a seguir:

- Abacate;
- Azeitona;
- Banana;
- Cacau;
- Café;
- Caqui;
- Castanha de caju;
- Dendê;
- Erva-mate;
- Figo;
- Goiaba;
- Laranja;
- Limão;
- Maçã;
- Mamão;
- Manga;
- Maracujá;
- Marmelo;
- Noz;
- Palmito;
- Pera;
- Pêssego;
- Pimenta do reino;
- Tangerina;
- Uva;
- Abacaxi;
- Alho;
- Amendoim;
- Arroz;
- Aveia;
- Batata-doce;
- Batata-inglesa;
- Cana-de-açúcar;
- Cebola;
- Ervilha;
- Feijão;
- Girassol;
- Mandioca;
- Melancia;
- Melão;
- Milho;
- Soja;
- Tomate;
- Trigo.

Tabela 5.1: Atributos explicativos do atributo-alvo

Atributo	**Descrição**	**Fonte**
vlr_pib_corrente_1000	PIB corrente anual, em milhares de reais	IBGE
vlr_pbf	Valor anual transferido pelo 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 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

Tabela 5.1: Continuação

Atributo	Descrição	Fonte
qtd_vinculos_formais	Estoque de vínculos empregatícios formais registrados no mês de dezembro	RAIS
qtd_supermercados	Quantidade de supermercados média 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 PNAE	Dados GOV
qtd_animais	Quantidade de animais produzidos durante o ano	IBGE
vlr_lavoura_1000	Valor da produção agrícola, em milhares de reais	IBGE
vlr_produzida_lavoura_ton	Produção agrícola, 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
perc_adesao_politica_saneamento	Percentual da população que reside em município com Política de Saneamento Básica	SNIS

É, finalmente, montado um conjunto de dados com os atributos mencionados na Tabela 5.1, com 439 instâncias, sendo estas as regiões do SUS. O atributo-alvo é a taxa

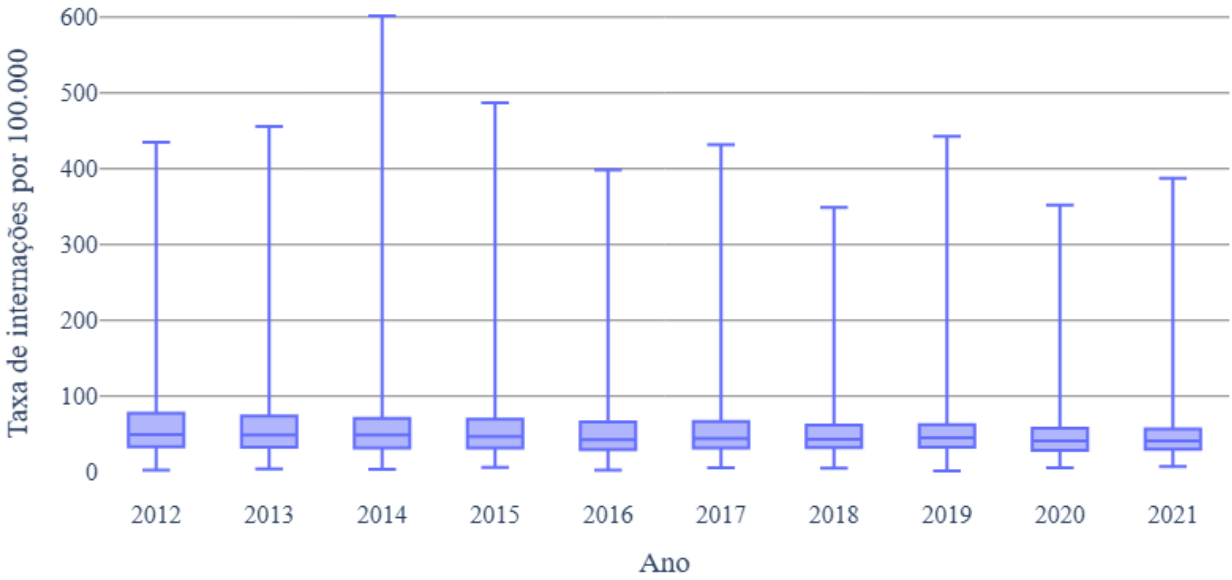
de internação anual por desnutrição por 100.000 pessoas. Os atributos explicativos são do período 2012-2021, e o atributo-alvo é do período 2013-2022, sendo que, para cada t, o atributo-alvo está em $t + 1$. Isto ocorre na intenção de se utilizar do modelo para prever o ano seguinte, para definir mudanças nas políticas públicas, se elas se fizerem necessárias.

As manipulações que resultaram no conjunto final foram feitas utilizando Knime, um programa *low code* de código aberto desenvolvido para a construção de *workflows* de tratamento e para a análise de dados.

5.1.1 Histórico da internação por desnutrição

O histórico de internação pela comorbidade de desnutrição é disponibilizado através do DATASUS, um portal de dados públicos atrelado ao Ministério da Saúde. Por meio deste canal, é possível acessar dados de internações e de procedimentos na rede pública de saúde de todo o país. São disponibilizados os dados no nível de município de residência, município de internação, idade, gênero e ano de internação. Neste trabalho, optou-se apenas pela segregação por ano e por município de residência.

Figura 5.1: Histórico da taxa de internação por desnutrição em t-1



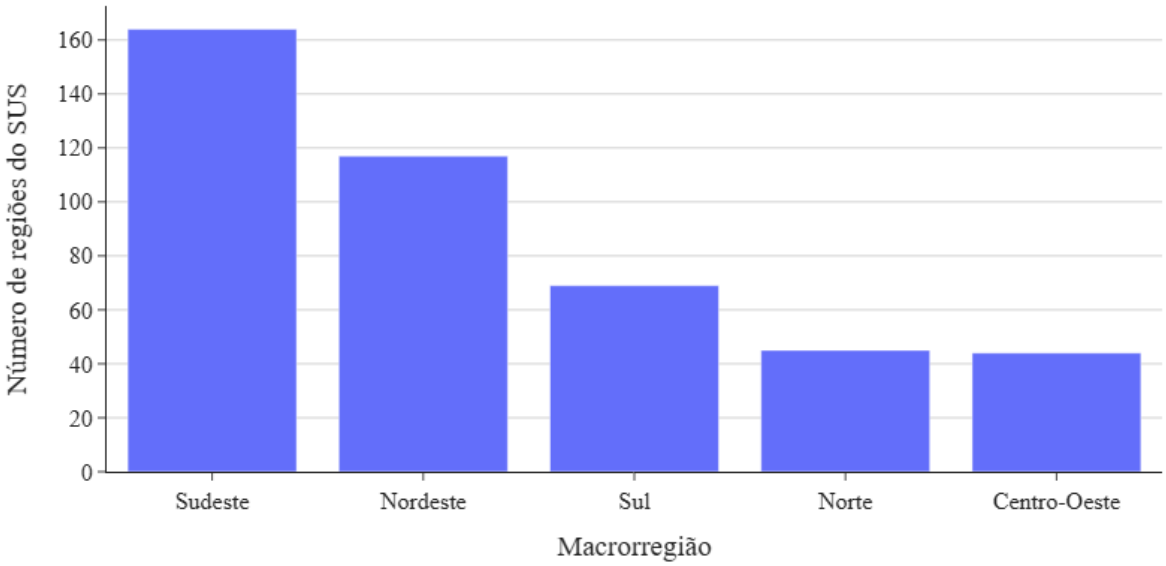
Fonte: De autoria própria.

Conforme mencionado anteriormente, o atributo-alvo é a taxa de internação por desnutrição por 100.000 pessoas e agregada nas regiões do SUS. A população de cada região é a fornecida pelo próprio DATASUS, optou-se por utilizar esse dado para manter-

se a coerência dentro da análise dos dados de saúde pública, uma vez que o número da população pode variar ligeiramente entre diferentes órgãos devido à distintos métodos para a estimativa. Na Figura 5.1, há o *boxplot* temporal do atributo-alvo, sem os *outliers*, os quais foram removidos pela regra do interquartil. É possível observar uma redução da mediana e do percentil 75% ao longo dos anos, o que corresponde a um raciocínio de indução conforme apresentado na literatura, onde se relaciona a melhora da taxa de internação à de indicadores sociais, conforme ocorreu nas últimas duas décadas no Brasil.

A taxa é agrupada dentro das regiões do SUS, as quais, por sua vez, são formadas por municípios fronteiriços que compartilham identidades culturais, econômicas e sociais, redes de comunicação e infraestrutura de transportes. Como a densidade populacional brasileira não é homogênea, é de se esperar que as regiões com maior população e maior quantidade de municípios também possuam uma quantidade maior de regiões do SUS, conforme apresentado na Figura 5.2. Tal comportamento se repete quando se analisam as unidades federativas individualmente, com os estados mais populosos apresentando mais regiões do SUS.

Figura 5.2: Distribuição das regiões do SUS nas macrorregiões do Brasil



Fonte: De autoria própria.

5.1.2 Análise exploratória dos atributos

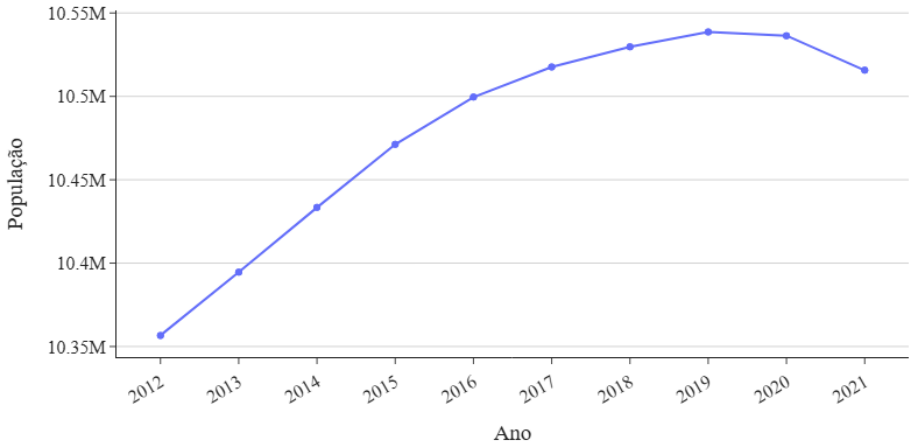
Os atributos foram coletados de diferentes órgãos governamentais, notadamente do IBGE. Os dados foram disponibilizados em formato CSV, contendo os dados abertos

por municipalidade e por ano. A partir disso, foi elaborado um *workflow* em Knime que unifica os atributos escolhidos dentro das Regiões do SUS de acordo com a população estimada do DATASUS.

A seguir, foi feita uma verificação da qualidade dos dados. Por serem dados governamentais, não há dados faltantes em nenhum município de cada ano. Em alguns casos específicos, novos municípios foram criados a partir do desmembramento de outros, porém, como há o agrupamento por Regiões do SUS, que foram fixadas de acordo com a listagem vigente (2025), não há a necessidade de descartar estes municípios nem de se realizar um *input* com dado sintético.

Também foi analisada a correlação de Pearson, métrica que avalia cada par de atributos, mensurando o quão fortemente correlacionados eles estão. Quando a métrica aproxima-se de 1 para dado par de atributos, significa que ambos movem-se na mesma direção e que se pode descartar um deles por ser redundante. A partir desta análise, foi descartado o atributo de Índice de Gini, pois ele apresentava alta correlação com vários outros atributos aqui considerados. É de se considerar que, por o Índice de Gini ser um valor que representa a desigualdade social, outros indicadores socio-econômicos possam abarcar o significado deste atributo.

Figura 5.3: População total

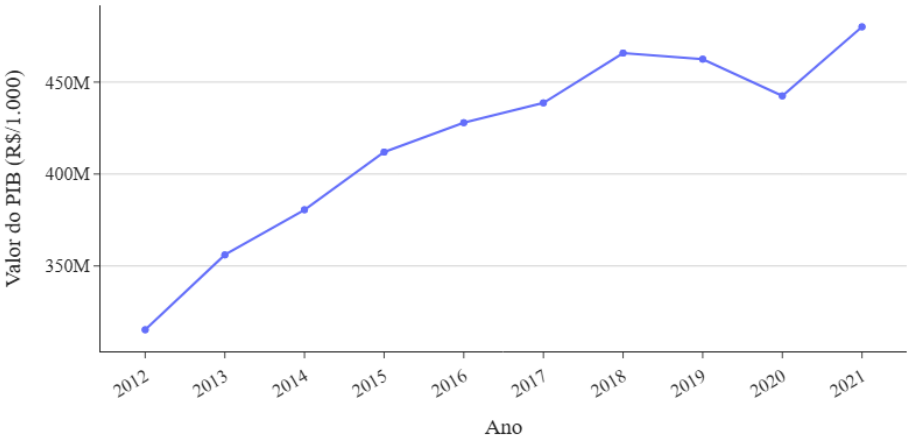


Fonte: De autoria própria.

As Figuras 5.3, 5.4, 5.5 e 5.6 apresentam exemplo de quatro atributos para a região 33005, região metropolitana da cidade do Rio de Janeiro e Região do SUS mais populosa. É visível nestes atributos o efeito da pandemia do Covid-19, por exemplo, na Figura 5.4, no ano 2020, há uma queda brusca no PIB, assim como em 5.6, no ano de 2021, há um aumento significativo das pessoas atendidas pelo Programa Bolsa Família, ano em que o programa foi reformulado como Auxílio Brasil e passou a abarcar mais pessoas devido à

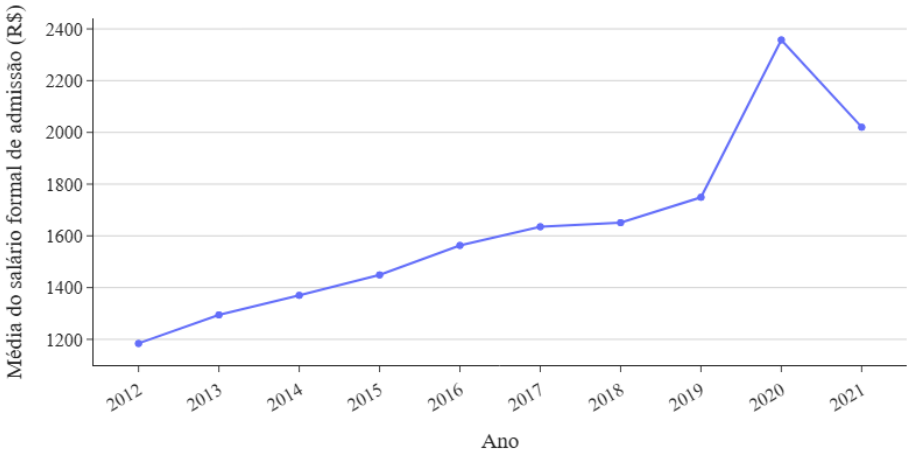
pandemia.

Figura 5.4: Valor do Produto Interno Bruto



Fonte: De autoria própria.

Figura 5.5: Valor médio do salário formal de admissão no mês de dezembro



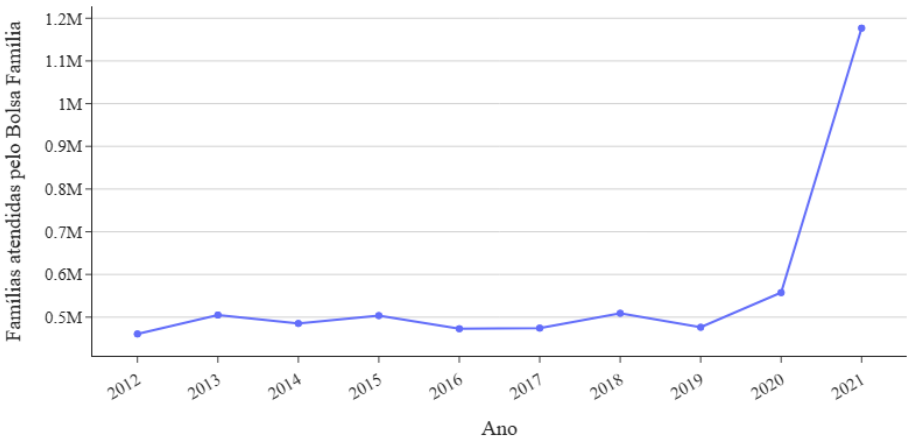
Fonte: De autoria própria.

5.2 Montagem e treinamento dos modelos

Todo o processo que envolve o pré-processamento, o treino e a análise dos modelos foi realizado em ambiente Python, utilizando-se o *VS Code* como interface de desenvolvimento. Para manipulação dos dados, foram utilizadas as bibliotecas Pandas e Numpy; para a criação de gráficos, as bibliotecas Matplotlib, *Seaborn* e Plotly. Já para os modelos, foram utilizadas LGBM, *Sklearn* e *XGBoost*, além das auxiliares Skopt, SHAP e *Optbinning*. Suas utilizações serão descritos nesta seção.

Para o *dataset*, foram treinados três modelos, LGBM, *XGBoost* e *Random Forest*,

Figura 5.6: Pessoas atendidas pelo Programa Bolsa Família



Fonte: De autoria própria.

porém, antes do treinamento em si, foi realizado um pré-processamento dos dados, descritos na seção a seguir. Também foi implementado um modelo *baseline* a fim de se ter uma base de comparação. O modelo *baseline* escolhido é uma simples regressão linear, treinada com todos os atributos-explicativos, sem *feature engineering*.

5.2.1 Pré-processamento para o treinamento

O conjunto de dados foi padronizado utilizando o método *StandardScaler* da biblioteca *Sklearn*, que redimensiona os atributos de acordo com a distribuição normal, fazendo com que a média seja zero e a variância seja unitária. 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 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.

O *dataset* não possui variáveis categóricas, então não é necessário realizar nenhuma transformação, como o *label encoding*.

Utilizou-se da função *Train_test_split*, também da biblioteca *Sklearn*, para se realizar a divisão dos dados para os conjuntos de treino, teste e OOT. Para a validação OOT, foi escolhido o último ano do conjunto, 2021. Já os anos anteriores, 2012 a 2020, 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.

A padronização foi realizada dentro do subconjunto de treino e depois aplicada para os subconjuntos de testes e validação, visando evitar *data leakage* no desenvolvi-

mento dos modelos.

5.2.2 Seleção de atributos

Para a identificação dos atributos que performam melhor com os modelos, foi realizada uma análise de Correlação de Pearson, conforme mencionado na seção anterior. Além disso, foi aplicada a técnica RFE, utilizando-se a função RFECV, também da biblioteca *Sklearn*, para a escolha de atributos. A técnica foi aplicada sobre os dados de treino padronizados.

Para cada modelo, foi utilizado como estimador o próprio modelo, por exemplo, para o modelo LGBM, foi utilizado como estimador o *LGBMRegressor*. Além disso, a métrica para escolha de atributos escolhida foi o R^2, com validação cruzada de cinco folhas e passo unitário, ou seja, a cada iteração, um atributo é removido. O número de atributos escolhidos é aquele que maximiza a métrica média da validação cruzada.

5.2.3 Otimização de hiperparâmetros

A seleção dos melhores hiperparâmetros para os modelos foi feita com a função *BayesSearchCV* da biblioteca Skopt, que implementa uma pesquisa com otimização bayesiana sobre os hiperparâmetros. Neste método, nem todas combinações de hiperparâmetros são testadas, apenas um número fixo que é amostrado de distribuições específicas. O número de combinações que é testada é dada por n_iter, cujo número padrão é cinquenta, porém, neste trabalho, optou-se por trinta e duas iterações, pois um número maior de iterações não resultava em melhora significativa da métrica e provoca um crescimento significativo do tempo para se encontrar os hiperparâmetros.

Para cada um dos modelos, foi dado um conjunto de hiperparâmetros a ser testado, de acordo com a especifidade de cada um, para a otimização nos sucessivos testes, buscando minimizar ou maximizar determinada métrica de avaliação. Para o trabalho, definiu-se como objetivo do otimizador a maximização do R^2. A função objetivo do *BayesSearchCV* realiza validação cruzada em seus testes, optando-se por cinco *folds* como argumento da função.

5.2.4 Categorização

Uma técnica comum de *feature engineering*, que visa melhorar a performance de modelos, é a categorização de variáveis numéricas. Essencialmente, significa agrupar variáveis numéricas, usualmente contínuas, em conjuntos distintos, podendo ser através da aplicação de conhecimento da área ou de algoritmos de categorização.

Alguns modelos podem se beneficiar da redução da granularidade da informação, especialmente modelos baseados em árvores, como o *Random Forest*. Assim, foi testada a categorização de alguns atributos, visando uma melhoria de performance, utilizando a biblioteca *OptBinning* para atributos contínuos. A categorização foi realizada sobre o subconjunto de treino e, então, aplicada aos subconjuntos de teste e OOT.

5.3 Explicabilidade

Técnicas que auxiliam na explicabilidade de modelos são importantes para facilitar a compreensão da importância de cada atributo e como este atributo contribui para o resultado final. No contexto deste trabalho, foi utilizado o SHAP como ferramenta para a explicabilidade, 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.

Um dos objetivos secundários deste trabalho é a compreensão da influência da segmentação territorial. Assim, optou-se por utilizar técnicas de redução de dimensionalidade, com o objetivo de entender se há diferença de comportamento dos atributos entre as macrorregiões do Brasil ou, ainda, de identificar se através dos atributos seria possível distinguir as macrorregiões.

Neste trabalho, foi implementada uma projeção PCA, utilizando-se da função PCA da biblioteca *Sklearn*, com dois componentes e uma projeção t-SNE, também utilizando-se da biblioteca *Sklearn* com dois componentes.

5.4 Implementação prática

A implementação dos modelos LGBM foi realizada utilizando-se da biblioteca *LightGBM*, a qual é uma solução proprietária da Microsoft e bastante popular no campo de aprendizado de máquina. Otimizaram-se os seguintes hiperparâmetros:

- *num_leaves*: máximo número de nós para as árvores base;
- *max_depth*: máxima profundidade das árvores base;
- *learning_rate*: taxa de aprendizado;
- *n_estimators*: número de árvores a ajustar;
- *subsample*: razão para subamostra do conjunto de teste;
- *colsample_bytree*: razão para subamostra de colunas quando construindo cada árvore;
- *min_child_samples*: número mínimo de dados necessário para um nó;
- *reg_alpha*: regularização L1;
- *reg_lambda*: regularização L2.

Os hiperparâmetros relacionados à regularização têm como objetivo principal evitar o *overfitting*, algo bastante comum em modelos com *boosting*, como é o caso do LGBM. Limitar a profundidade do modelo e o número máximo de nós também auxilia na prevenção do *overfitting*, sendo uma escolha delicada, pois limitar excessivamente os parâmetros pode prejudicar o aprendizado, levando a *underfitting*.

Já os modelos *Random Forest* foram implementados utilizando-se da biblioteca *Sklearn*. Otimizaram-se os seguintes hiperparâmetros:

- *max_depth*: máxima profundidade das árvores base;
- *n_estimators*: número de árvores a ajustar;
- *min_samples_split*: número mínimo de amostras para dividir um nó;
- *min_samples_leaf*: número mínimo de amostras para um nó que não tem filhos;
- *max_features*: número máximo de atributos em cada divisão.

Novamente, a limitação de profundidade e de atributos ajuda a limitar o crescimento das árvores, evitando o *overfitting*. É interessante, também, limitar o crescimento para reduzir o consumo de memória.

Os modelos *XGBoost* foram implementados utilizando-se a biblioteca *XGBoost*, com a otimização dos hiperparâmetros a seguir:

- *n_estimators*: número de árvores a ajustar;
- *max_depth*: máxima profundidade das árvores base;
- *learning_rate*: taxa de aprendizado;
- *subsample*: razão para subamostra do conjunto de teste;
- *colsample_bytree*: razão para subamostra de colunas quando construindo cada árvore;
- *reg_alpha*: regularização L1;
- *reg_lambda*: regularização L2.

A Tabela 5.2 resume as matrizes de hiperparâmetros testadas com o algoritmo de pesquisa bayesiana. Incluem-se, para os hiperparâmetros numéricos, os valores mínimos e máximos registrados, além da função de geração dos números.

Tabela 5.2: Hiperparâmetros para LGBM, *Random Forest* e *XGBoost*

Algoritmo	**Hiperparâmetro**	**Faixa de valores**
LGBM	*num_leaves*	[2,128]
	max_depth	[1,10]
	learning_rate	[0.01,0.1] (*log-uniform*)
	n_estimators	[100,1000]
	subsample	[0.5,1.0] (*uniform*)
	colsample_bytree	[0.5,1.0] (*uniform*)
	reg_alpha	[1e-9,100] (*log-uniform*)
	reg_lambda	[30,100] (*log-uniform*)
	min_child_samples	[5,30]
Random Forest	*n_estimators*	[100,1000]
	max_depth	[2,30]
	min_samples_split	[2,20]
	min_samples_leaf	[1,20]
	max_features	[0.3,1.0] (*uniform*)
XGBoost	*n_estimators*	[100,1000]
	max_depth	[2,20]
	learning_rate	[0.01,0.5] (*log-uniform*)
	subsample	[0.5,1.0] (*uniform*)
	colsample_bytree	[0.5,1.0] (*uniform*)
	reg_alpha	[1e-9,100] (*log-uniform*)
	reg_lambda	[1e-9,100] (*log-uniform*)

Além das implementações mencionadas nesta seção, também se realizaram testes, adicionando-se como atributos a média de cada atributo nos dois anos anteriores do atributo-alvo. Por exemplo, para a predição em um ano t, utilizou-se a média de cada atributo em $t-1$ e $t-2$, invés de utilizar apenas do valor do atributo em $t-1$. O objetivo é observar se a mudança de políticas que afetem os atributos demoram a refletir na redução

de internações por desnutrição, de modo que utilizar uma média de um tempo anterior reflita de maneira mais precisa o comportamento do que apenas o ano anterior.

Por fim, a implementação prática do trabalho segue uma sequência estruturada, aplicada de forma consistente aos modelos LGBM, *Random Forest* e *XGBoost*. Inicialmente, procede-se à padronização dos atributos. Após essa preparação, realiza-se a seleção de atributos por meio do método RFE, permitindo a identificação dos atributos mais relevantes para cada modelo. Sobre o conjunto de atributos previamente selecionados e padronizados, aplica-se a busca por hiperparâmetros.

Na etapa seguinte, incorporam-se as médias dos dois anos anteriores a cada instante t, repetindo-se o processo de seleção de atributos e otimização de hiperparâmetros para avaliar o impacto dessa informação temporal adicional. Nesta etapa, o período temporal utilizado para o atributo-alvo é entre 2014 e 2022, sendo o ano de 2022 utilizado para validação OOT. Posteriormente, também é implementada a categorização dos atributos utilizando-se a função *ContinuousOptimalPWBinning* da biblioteca *OptBinning*, cujo desempenho é analisado em comparação às abordagens anteriores.

Após, para cada macrorregião do país, manteve-se as médias dos dois anos anteriores a cada t, aplicando-se novamente a seleção de atributos e a busca pelos melhores hiperparâmetros, de modo a capturar possíveis heterogeneidades regionais, assim, foram obtidos três modelos, *Random Forest*, LGBM e *XGBoost* para cada macrorregião do Brasil. Comparou-se também os resultados destes modelos regionais com os modelos globais construídos anteriormente.

Por fim, empregam-se técnicas de redução de dimensionalidade, especificamente PCA e t-SNE, com o propósito de auxiliar na interpretação da influência dos atributos tanto nos modelos regionais quanto no modelo global, contribuindo para uma compreensão mais aprofundada da estrutura dos dados e do comportamento dos algoritmos.

6 RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados os resultados obtidos a partir do treinamento, validação e avaliação dos modelos preditivos desenvolvidos para estimar a taxa de internação por desnutrição nas 439 regiões de saúde do SUS. Os resultados incluem desempenho dos modelos, seleção de atributos, explicabilidade e avaliação OOT.

6.1 Modelo *baseline*

A fim de se comparar os resultados dos modelos implementados neste trabalho, é treinado uma regressão linear, sem *feature engineering*, a fim de se ter um modelo *baseline*. As métricas são apresentadas na Tabela 6.1. Observa-se que uma simples regressão linear não captura o comportamento em relação ao atributo alvo, uma vez que o coeficiente de determinação é praticamente nulo e o MAPE é próximo de um.

Tabela 6.1: Métricas do modelo *baseline* baseado em regressão linear

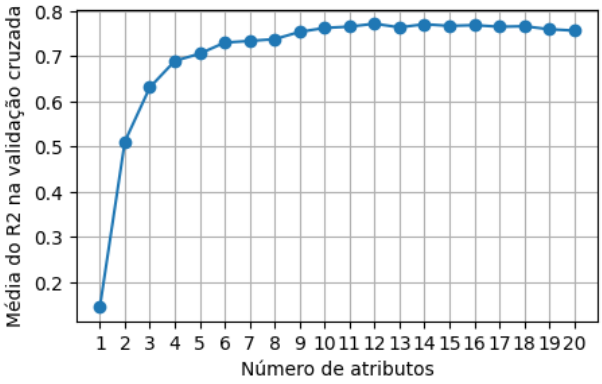
	Treino	**Teste**	**OOT**
R^2	0,08	0,07	-0,06
MAPE	0,79	0,74	0,59

6.2 Seleção de atributos

Após a padronização dos atributos com o método *StandardScaler*, foi realizada a seleção com a técnica RFE para cada modelo, a qual é apresentada nas Figuras 6.1, 6.2 e 6.3. Nelas, podemos observar que 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.

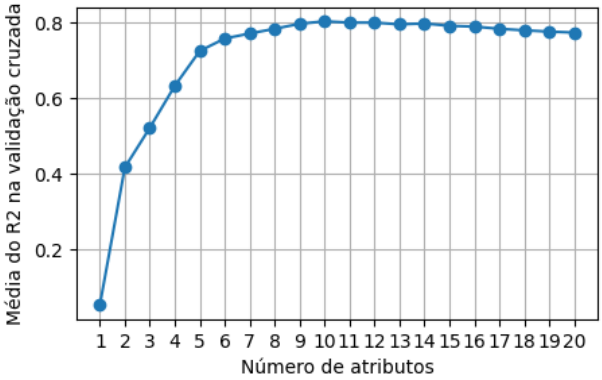
Para o modelo LGBM, o melhor número de atributos é doze; para o *Random Forest*, é dez; e para o *XGBoost*, é quinze. Isso é mostrado na Tabela 6.2, onde podemos observar que a maioria dos atributos se repete para os três modelos. Na tabela, pode-se ver que alguns atributos não foram selecionados para nenhum modelo, sendo eles o vlr_bpc (valor anual do Benefício de Prestação Continuada), vlr_medio_salario_formal (valor médio do salário formal de admissão no mês de dezembro), qtd_supermercados

Figura 6.1: Seleção de atributos usando RFE com LGBM



Fonte: De autoria própria.

Figura 6.2: Seleção de atributos usando RFE com *Random Forest*

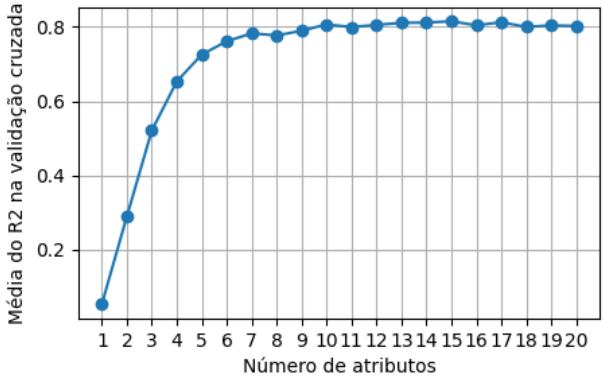


Fonte: De autoria própria.

(quantidade de supermercados), pop_total_atendida_agua (população total atendida com abastecimento de água tratada), pop_total_atendida_coleta (população total atendida com coleta de resíduos) e perc_adesao_politica_saneamento (percentual da população em municípios com adesão à Política de Saneamento Básica). 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 mostrou relevante. Também, mostra-se importante o atributo qtd_alunos_pnae, que indica quantos estudantes são atendidos pelo Programa Nacional de Alimentação Escolar, 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).

Assim como no estudo de Silva et al. (2025), o saneamento básico se mostrou como atributo relevante, uma vez que pop_total_atendida_esgoto foi selecionado para os três modelos. É interessante notar que, embora a quantidade de supermercados não tenha sido um atributo selecionado por nenhum modelo, a produção de animais e de alimentos foram atributos selecionados, o que, no Brasil, pode representar um comportamento e/ou

Figura 6.3: Seleção de atributos usando RFE com *XGBoost*



Fonte: De autoria própria.

cultura diferente de países como os Estados Unidos, por exemplo, onde a quantidade de supermercados e a sua localização representam um fator importante na qualidade da alimentação, levando inclusive à existência de ações governamentais para mitigar esta questão (HUA et al., 2024).

Tabela 6.2: Atributos selecionados para cada modelo

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000	x	x	x
vlr_pbf			x
qtd_familias_pbf	x	x	x
vlr_bpc			
qtd_beneficiarios_bpc	x	x	x
vlr_medio_salario_formal			
qtd_vinculos_formais	x	x	x
qtd_supermercados			
perc_mulheres	x	x	x
qtd_populacao	x	x	x
qtd_escolas	x	x	x
qtd_alunos_ensino_basico			x
qtd_alunos_pnae	x		x
qtd_animais	x	x	x
vlr_lavoura_1000	x		x
qtd_produzida_lavoura_ton	x	x	x
pop_total_atendida_agua			x
pop_total_atendida_esgoto	x	x	x
pop_total_atendida_coleta			
perc_adesao_politica_saneamento			

A Tabela 6.3 mostra as métricas após a padronização dos atributos e a seleção para cada modelo. Podemos observar que todos os modelos têm uma piora em ambas as métricas em OOT, embora o modelo *XGBoost* tenha tido a queda mais expressiva, o que

pode indicar *overfitting*. Nesta etapa, os modelos foram treinados com os hiperparâmetros padrão de cada biblioteca.

Tabela 6.3: Métricas dos modelos após seleção de atributos

Métrica	**R^2**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,93	0,98	0,99
Teste	0,83	0,85	0,87
OOT	0,54	0,48	0,35
Métrica	**MAPE**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,22	0,08	0,08
Teste	0,29	0,22	0,21
OOT	0,40	0,41	0,49

6.3 Otimização bayesiana

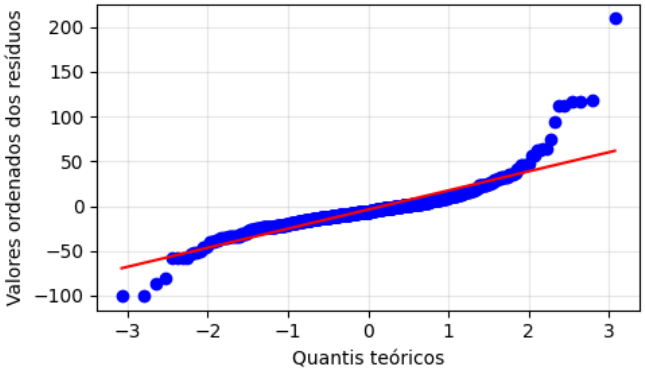
Após a padronização e seleção de atributos, foi aplicada a otimização de hiperparâmetros com a técnica de otimização bayesiana, na matriz apresentada na Tabela 5.2, com validação cruzada de cinco *folds*, trinta e duas iterações e otimizando a métrica R^2. Os resultados são apresentados na Tabela 6.4 abaixo.

Tabela 6.4: Métricas dos modelos após seleção de atributos e busca por hiperparâmetros

Métrica	**R^2**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,99	0,98	0,99
Teste	0,90	0,88	0,90
OOT	0,56	0,53	0,58
Métrica	**MAPE**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,04	0,08	0,03
Teste	0,16	0,20	0,15
OOT	0,38	0,44	0,39

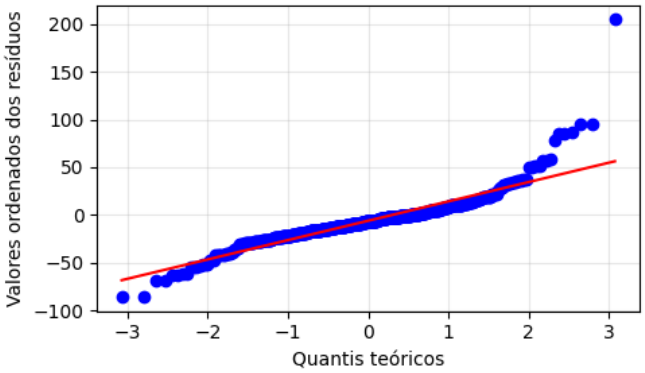
Ao comparar os resultados apresentados nas Tabelas 6.3 e 6.4, verifica-se uma melhoria geral das métricas de desempenho, com destaque para o modelo *XGBoost*, cujo MAPE apresenta uma redução de aproximadamente 10%. No caso do modelo *Random Forest*, observa-se um incremento no valor do R^2, acompanhado, entretanto, de uma discreta piora no MAPE. É também verdade que, usualmente, modelos *Random Forest* não melhoram significativamente com a otimização de hiperparâmetros, uma vez que, por natureza, são modelos mais robustos e de menor dependência de ajustes finos.

Figura 6.4: Q-Q para *Random Forest*



Fonte: De autoria própria.

Figura 6.5: Q-Q para *Random Forest* com otimização



Fonte: De autoria própria.

A análise do gráfico *quantile-quantile* (Q–Q), referente aos resíduos no subconjunto OOT, evidencia que o processo de otimização conduzido pela função *BayesSearchCV*, ao priorizar a maximização do R^2, tende a reduzir os resíduos de maior magnitude absoluta. Contudo, essa otimização não assegura a preservação do comportamento dos resíduos de menor magnitude, o que pode resultar em penalizações na métrica MAPE. Como o MAPE, é sensível a variações percentuais, discrepâncias relativamente pequenas em termos absolutos podem representar diferenças percentuais expressivas, impactando negativamente o valor final da métrica.

6.4 Adição de atributos

Nesta etapa, substituiu-se os atributos pelas médias dos dois anos anteriores a cada ano t para cada atributo, repetindo-se o processo de seleção, utilizando RFE, e de

otimização de hiperparâmetros, utilizando *BayesSearchCV*, para avaliar o impacto dessa informação temporal adicional. Obtendo-se, então, os atributos selecionados apresentados na Tabela 6.5.

Tabela 6.5: Atributos selecionados para cada modelo após adição de variáveis

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x	x	x
vlr_pbf_media_2anos			
qtd_familias_pbf_media_2anos	x	x	x
vlr_bpc_media_2anos			
qtd_beneficiarios_bpc_media_2anos	x	x	x
vlr_medio_salario_formal_media_2anos			
qtd_vinculos_formais_media_2anos	x	x	x
qtd_supermercados_media_2anos			x
perc_mulheres_media_2anos	x	x	x
qtd_populacao_media_2anos	x	x	x
qtd_escolas_media_2anos	x	x	x
qtd_alunos_ensino_basico_media_2anos		x	x
qtd_alunos_pnae_media_2anos	x		x
qtd_animais_media_2anos	x	x	x
vlr_lavoura_1000_media_2anos	x	x	x
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos			
pop_total_atendida_esgoto_media_2anos	x	x	x
pop_total_atendida_coleta_media_2anos			
perc_adesao_politica_saneamento_media_2anos			

A comparação entre as Tabelas 6.2 e 6.5 mostra que, embora a inclusão das médias dos dois anos anteriores não tenha alterado de forma radical o conjunto de atributos considerados relevantes, ela modificou a quantidade total de variáveis selecionadas por cada modelo, indicando mudanças na sensibilidade temporal das predições. O modelo LGBM manteve o número de atributos selecionados em doze; o *Random Forest* passou de dez para doze; e o *XGBoost* de quinze para catorze. Na Tabela 6.5, observa-se que a mudança no número de atributos escolhidos sugere que a informação temporal suavizada ampliou a capacidade explicativa de atributos antes descartados, como qtd_supermercados e qtd_alunos_ensino_basico. Apesar disso, alguns atributos ainda permanecem em todas as situações, como o vlr_pib_corrente_1000, qtd_familias_pbf e qtd_vinculos_formais, enquanto atributos como vlr_pbf, vlr_bpc, vlr_medio_salario_formal e indicadores de abastecimento de água tratada e coleta de resíduos continuam sendo consistentemente excluídos, reforçando que, mesmo com mais informação temporal, eles não contribuem de forma significativa para a predição da taxa de internação por desnutrição nestes modelos

analisados.

Tabela 6.6: Métricas dos modelos após adição de atributos e busca por hiperparâmetros

Métrica	**R^2**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,99	0,98	0,99
Teste	0,83	0,80	0,85
OOT	0,66	0,64	0,65
Métrica	**MAPE**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,06	0,09	0,01
Teste	0,22	0,35	0,21
OOT	0,35	0,42	0,37

Os resultados dos modelos são apresentados na Tabela 6.6, onde podemos observar que as métricas melhoram em relação ao que é apresentado na Tabela 6.4. O R^2 melhora de mais significativa, especialmente no modelo *Random Forest*, embora, por ambas as métricas, o melhor desempenho seja do modelo LGBM. É possível observar também que há um claro ganho em relação ao modelo *baseline*, cujas métricas foram apresentadas na Tabela 6.1.

6.5 Categorização de atributos contínuos

Realizou-se um experimento utilizando a categorização dos atributos, como técnica de engenharia de *feature*, buscando melhorar as métricas. Conforme mencionado na Seção 5.4, utilizou-se da função *ContinuousOptimalPWBinning*, que aplica métodos de otimização para definir automaticamente os intervalos, buscando maximizar a relação entre os atributos categorizados e o atributo-alvo, com um mínimo de duas e um máximo de cinco categorias para cada atributo. Não se categorizou os atributos que são percentuais (perc_mulheres e perc_adesao_politica_saneamento) e o valor médio do salário formal de admissão. Aplicou-se a categorização sobre os atributos com as médias dos dois anos anteriores a cada ano t e, para cada modelo, foram utilizados os atributos selecionados pelo método RFE, apresentados na Tabela 6.5. A categorização foi aplicada sobre o subconjunto de treino. Após a categorização, os modelos foram novamente treinados e submetidos ao processo de otimização de hiperparâmetros, mantendo-se a mesma estratégia de validação utilizada nas etapas anteriores. Os resultados são apresentados na Tabela 6.7.

É interessante apontar que a categorização não melhora nenhuma das duas métri-

Tabela 6.7: Métricas dos modelos após categorização e busca por hiperparâmetros

Métrica	**R^2**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,95	0,97	0,98
Teste	0,78	0,75	0,79
OOT	0,47	0,49	0,51
Métrica	**MAPE**		
Conjunto/Modelo	**LGBM**	***Random Forest***	***XGBoost***
Treino	0,15	0,10	0,27
Teste	0,27	0,25	0,50
OOT	0,44	0,43	0,51

cas em nenhum modelo. No caso específico do *Random Forest*, a métrica MAPE praticamente não muda, embora fosse esperado que ela melhorasse, tendo em vista que a categorização, particularmente, pode aliviar o problema de repetição em árvores. Isso ocorre quando um mesmo atributo é testado repetidamente ao longo de um caminho em uma árvore de decisão, o que pode levar a uma complexidade desnecessária e a uma fragmentação dos dados (LIU; MOTODA, 2012). Os resultados apontam que a categorização não foi suficiente para compensar a redução da granularidade, assim, os resultados das seções subsequentes consideram os modelos treinados com atributos contínuos e médias temporais.

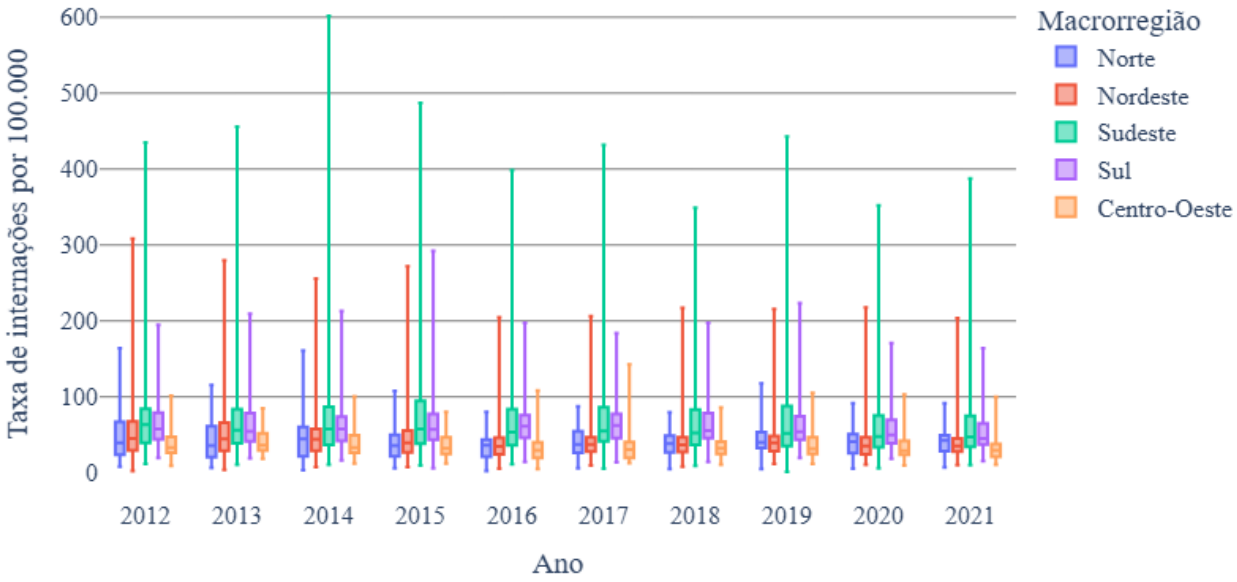
6.6 Desempenho dos modelos nas macrorregiões do país

Considerando os modelos treinados com os atributos médios dos dois anos anteriores, a cada ano t selecionados pelo algoritmo RFE e submetidos à busca por hiperparâmetros, segregou-se as métricas por cada macrorregião 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.

Na Figura 6.6, tem-se o *boxplot* da taxa de internação aberta por macrorregião. Pode-se observar que as macrorregiões Sul, Sudeste e Norte, em 2021, têm medianas muito similares, seguido das macrorregiões Nordeste e Centro-Oeste. Este comportamento varia ao longo da linha temporal, além disso as macrorregiões Sudeste, Sul e Nordeste possuem muita variabilidade em seus dados. Já na Figura 6.7, temos a média da taxa de internação por macrorregião, onde se observa que o Sul e Sudeste possuem, consistentemente, as maiores médias, seguidas do Norte e Nordeste, que alternam ao longo

do tempo a terceira e quarta posição. Por último, tem-se o Centro-Oeste.

Figura 6.6: *Boxplot* da taxa de internação por macrorregião (em t-1)



Fonte: De autoria própria.

Os resultados das avaliações para as macrorregiões estão na Tabela 6.8. 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.

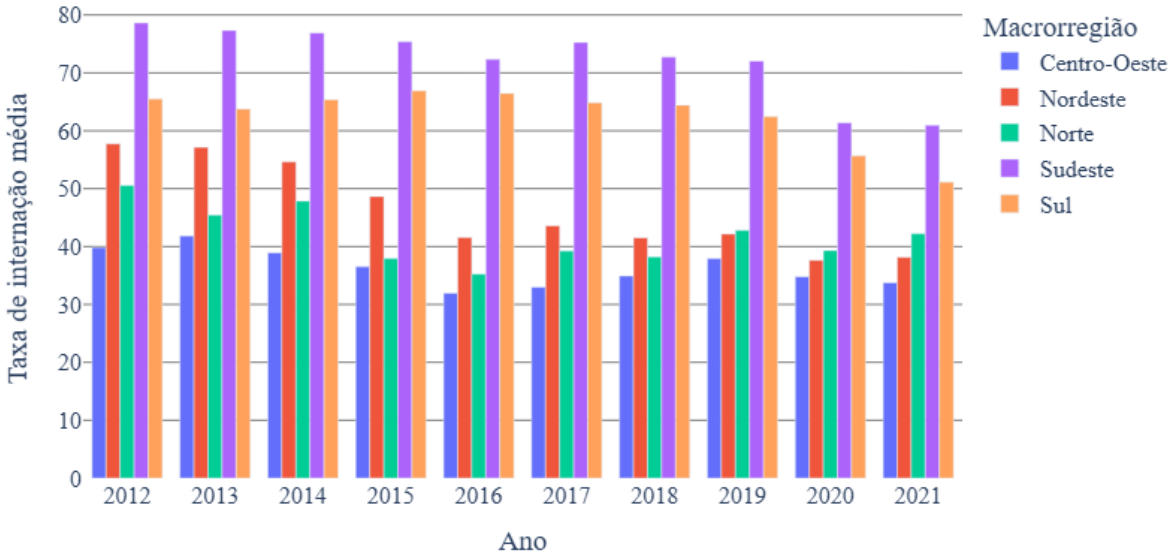
Considerando especificamente o comportamento observado para a macrorregião Centro-Oeste, os resultados apresentados na Tabela 6.8 indicam valores de R^2 muito baixos, próximos de zero, enquanto o MAPE se mantém em um patamar intermediário. Diante de um R^2 tão reduzido, seria esperado um valor de MAPE mais elevado, próximo de um; contudo, esse comportamento não se verifica em função da distribuição dos resíduos observada nessa região.

Tabela 6.8: Métricas da Tabela 6.6 abertas por macrorregião

Modelo	**LGBM**		***Random Forest***		***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

Conforme ilustrado na Figura 5.2, o Centro-Oeste é a macrorregião com o menor

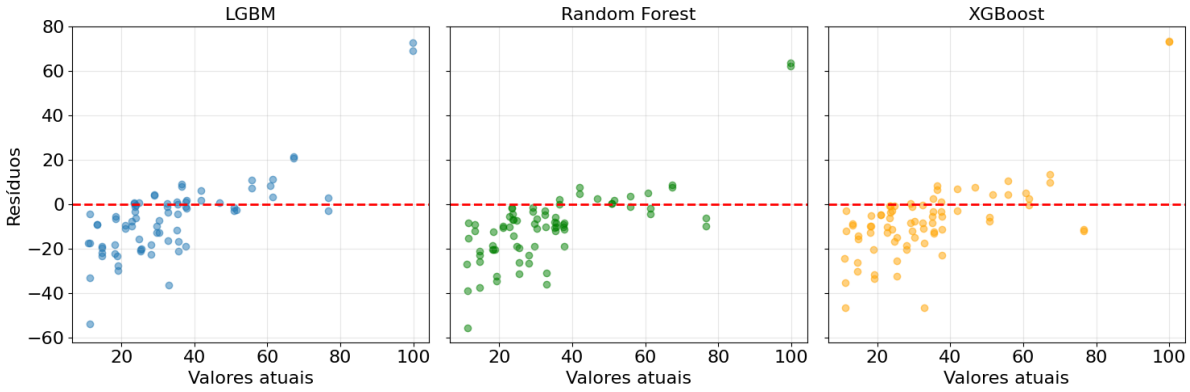
Figura 6.7: Taxa média de internação por macrorregião



Fonte: De autoria própria.

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 da taxa de internação por desnutrição nessa região, resultando em valores de R^2 muito baixos, conforme evidenciado na Figura 6.8. 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.

Figura 6.8: Resíduos vs. valores atuais no conjunto OOT para Centro-Oeste

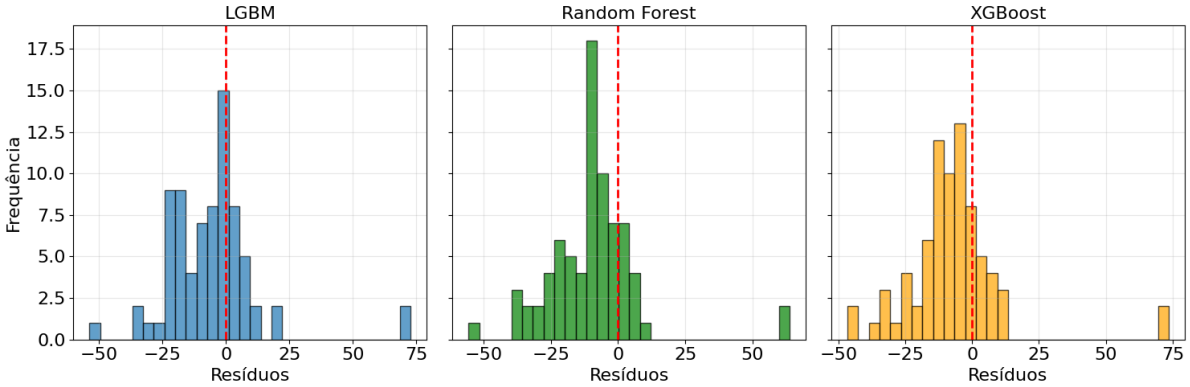


Fonte: De autoria própria.

Como se pode observar nas Figuras 6.9 e 6.10, é claro 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,

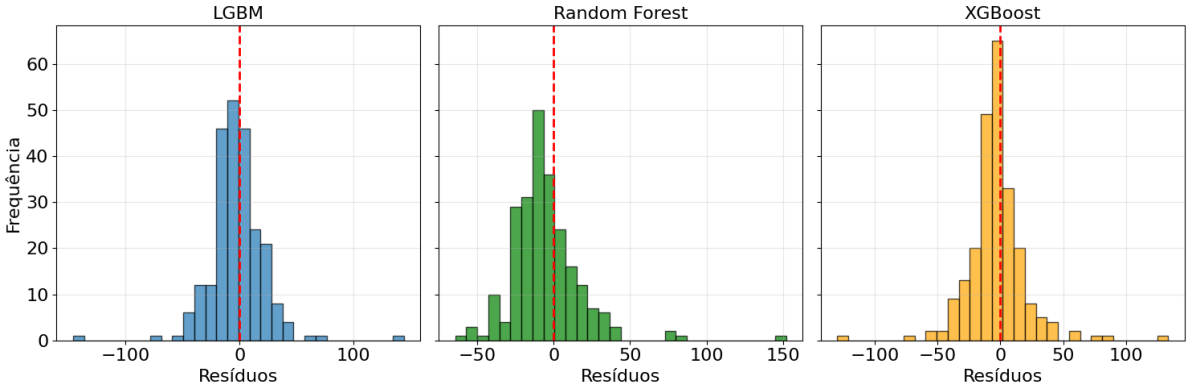
mostrando uma dispersão maior de resíduos. Já no Sul, conforme a Figura 6.11, há uma certa dispersão dos resíduos, porém o valor absoluto dos resíduos são menores o que leva a um MAPE menor.

Figura 6.9: Histograma de resíduos no conjunto OOT para Centro-Oeste



Fonte: De autoria própria.

Figura 6.10: Histograma de resíduos no conjunto OOT para Sudeste

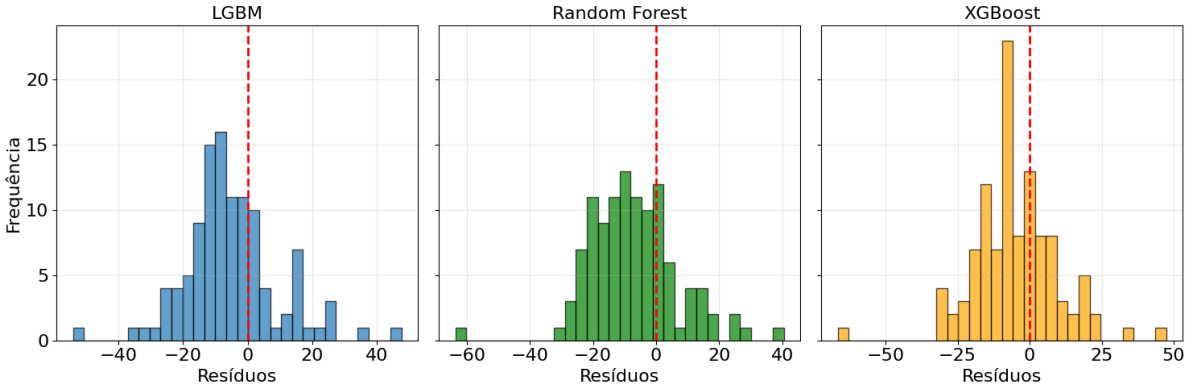


Fonte: De autoria própria.

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 região, os atributos utilizados são particularmente informativos e que o padrão das internações por desnutrição é mais estável e previsível. Isso possivelmente ocorre pelo fato de a 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.

Considerando os modelos treinados com os atributos médios dos dois anos anteriores, selecionados pelo algoritmo RFE e submetidos à busca por hiperparâmetros,

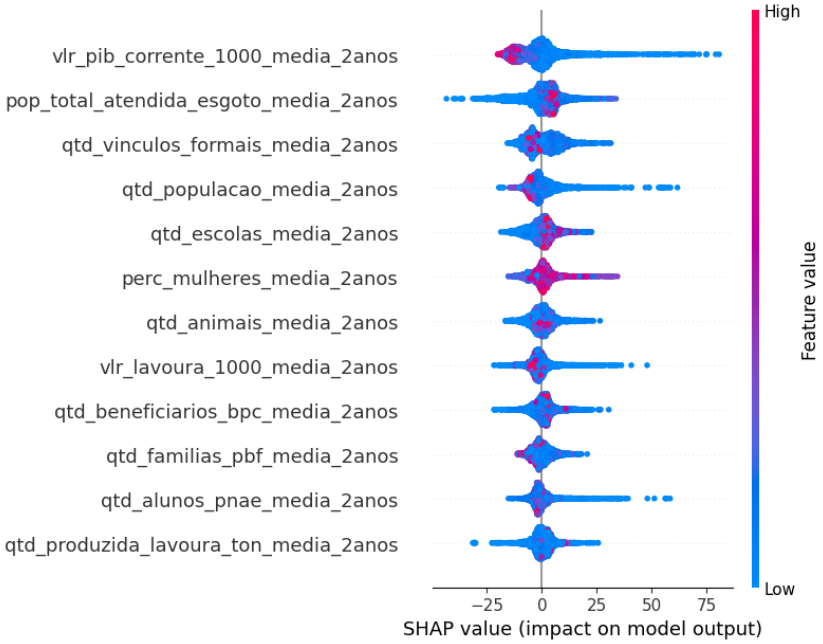
Figura 6.11: Histograma de resíduos no conjunto OOT para Sul



Fonte: De autoria própria.

escolheu-se o modelo LGBM, que apresentava as melhores métricas para o Brasil para se analisar o gráfico SHAP, o qual é uma ferramenta para auxiliar na explicabilidade. O SHAP apresenta uma ordem de importância dos atributos no modelo e ajuda a entender como a variação dos valores dos atributos pode alterar a saída do modelo. A ordenação vertical dos atributos reflete sua importância relativa no modelo, enquanto a dispersão dos pontos ao longo do eixo horizontal indica a magnitude e a direção do impacto no valor predito.

Figura 6.12: SHAP para Brasil (LGBM)

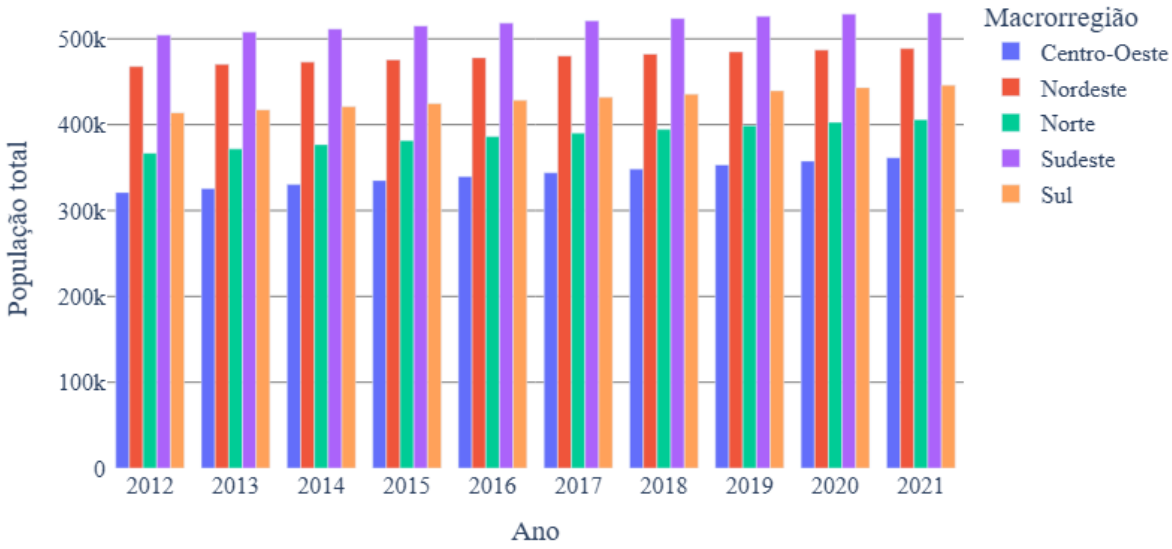


Fonte: De autoria própria.

A Figura 6.12 apresenta os gráficos de valores SHAP para o Brasil. 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 terceiro 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, assim, uma maior formalização do emprego pode levar à aumento da renda mensal. 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, como podemos ver nas Figuras 6.13 e 6.14.

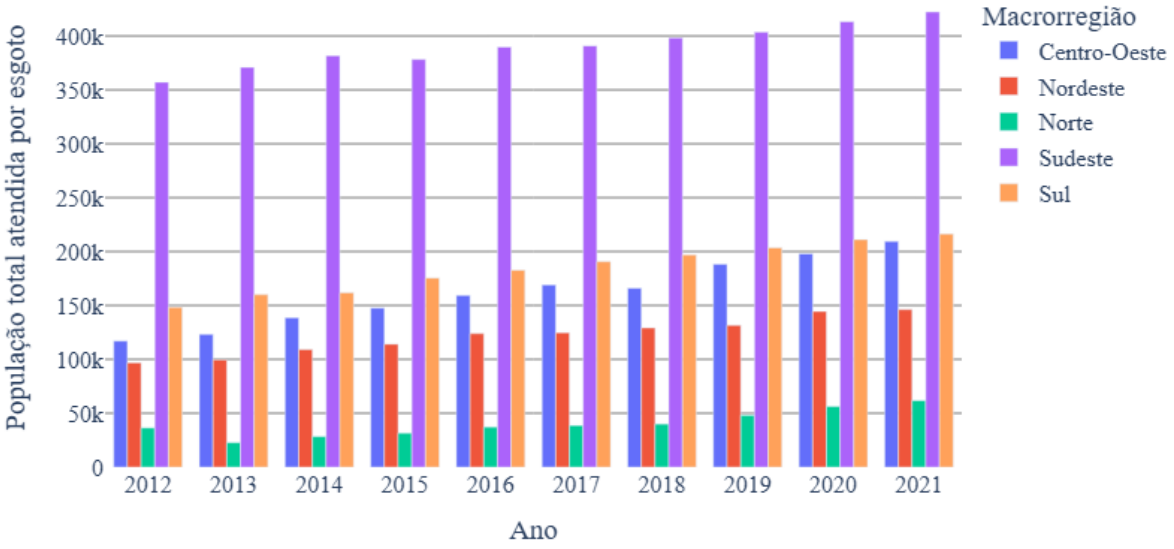
Figura 6.13: População total distribuída por macrorregião



Fonte: De autoria própria.

As Figuras 6.15 e 6.16 apresentam os gráficos de valores SHAP para as macrorregiões Sudeste e Sul, respectivamente. Para estas duas macrorregiões, observa-se que o PIB é o atributo de maior influência no modelo, assim como acontece quando se analisa o Brasil inteiro. É provável que este comportamento similar ao Brasil venha da quantidade de regiões de saúde nessas macrorregiões, que é bem superior às outras. Destacam-se

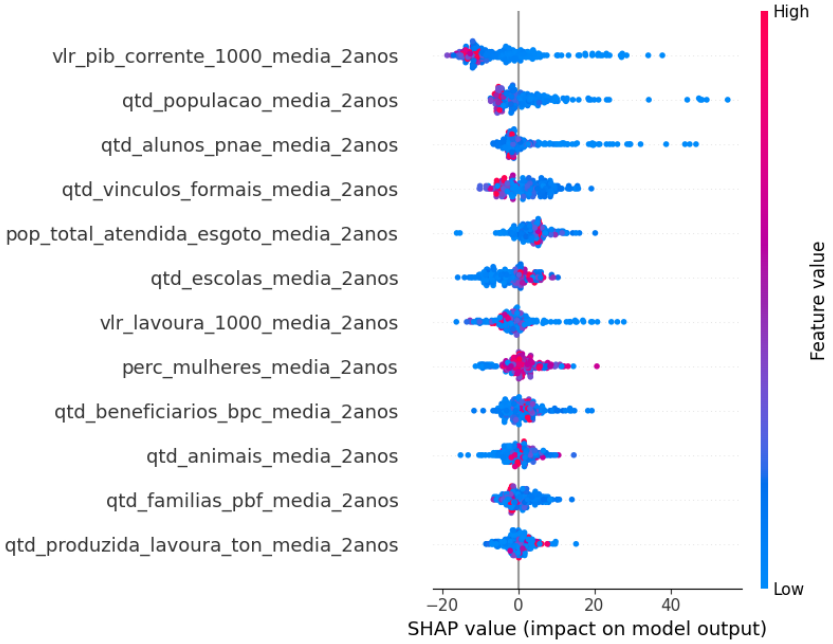
Figura 6.14: População total atendida por esgoto distribuída por macrorregião



Fonte: De autoria própria.

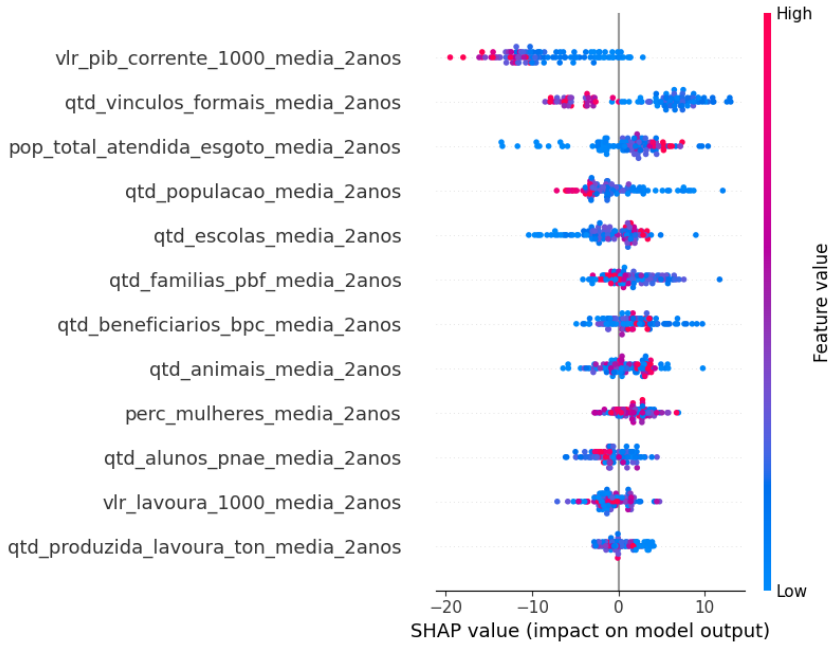
ainda a quantidade de famílias atendidas pelo PBF e o de alunos atendidos pelo PNAE, pois, apesar de apresentarem uma maior dispersão dos valores SHAP, é possível ver que quanto maior a quantidade de beneficiários destes programas sociais, menor a taxa de internação. Assim, estes programas sociais podem combater a vulnerabilidade social, que, eventualmente, vai levar à insegurança alimentar e à desnutrição.

Figura 6.15: SHAP para Sudeste (LGBM)



Fonte: De autoria própria.

Figura 6.16: SHAP para Sul (LGBM)



Fonte: De autoria própria.

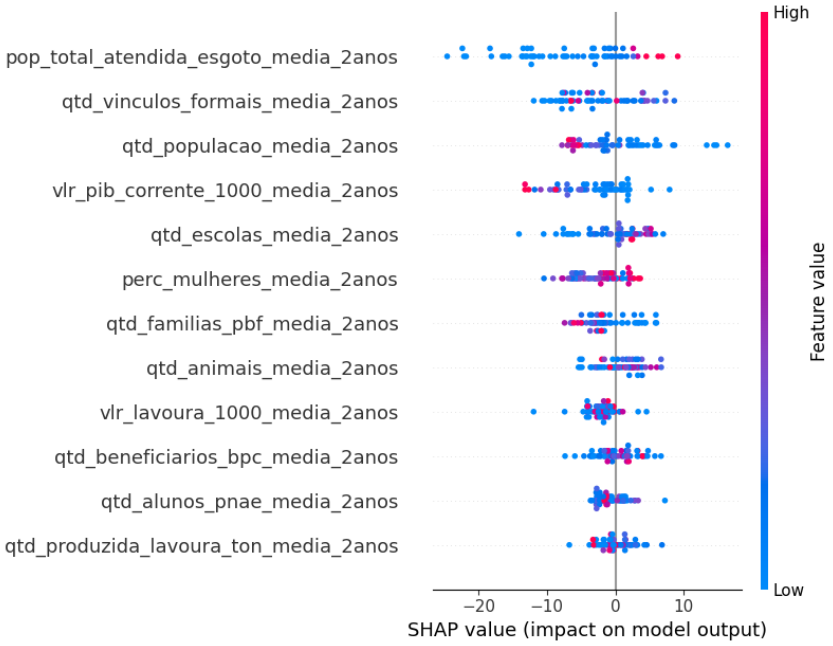
As Figuras 6.17 e 6.18 apresentam os gráficos de valores SHAP para as macrorregiões Norte e Nordeste, respectivamente. Para a macrorregião Nordeste, no atributo quantidade de vínculos formais, ainda é possível ver que quanto mais alto menor a taxa de internação, embora haja bastante dispersão dos valores, assim como na macrorregião Norte. E, embora em uma ordem de importância menor nestas macrorregiões, é claro que quanto maior o PIB menor a taxa de internação, assim como ainda é clara a importância do PBF como ferramenta de mitigação da vulnerabilidade social. Assim como acontece na análise Brasil, nas macrorregiões Norte e Nordeste, o atributo relacionado à população atendida com esgoto tratado aparece como uma variável de maior importância.

A Figura 6.19 apresenta o gráfico SHAP para a região Centro-Oeste, a qual é a região com as piores métricas, tanto R^2 quanto MAPE. Ainda que o modelo apresente desempenho inferior para a macrorregião Centro-Oeste, observa-se um padrão consistente com as demais macrorregiões, no qual o PIB figura como o atributo de maior importância. Nesse contexto, valores mais elevados de PIB tendem a estar associados a impactos negativos na predição da taxa de internação.

Os atributos que, consistentemente aparecem em todas as macrorregiões, são o valor do PIB, a quantidade de vínculos formais, a população total, a quantidade de alunos atendidos pelo PNAE, a população atendida pelo saneamento básico e a quantidade de famílias atendidas pelo Programa Bolsa Família. Especificamente este último atributo e a

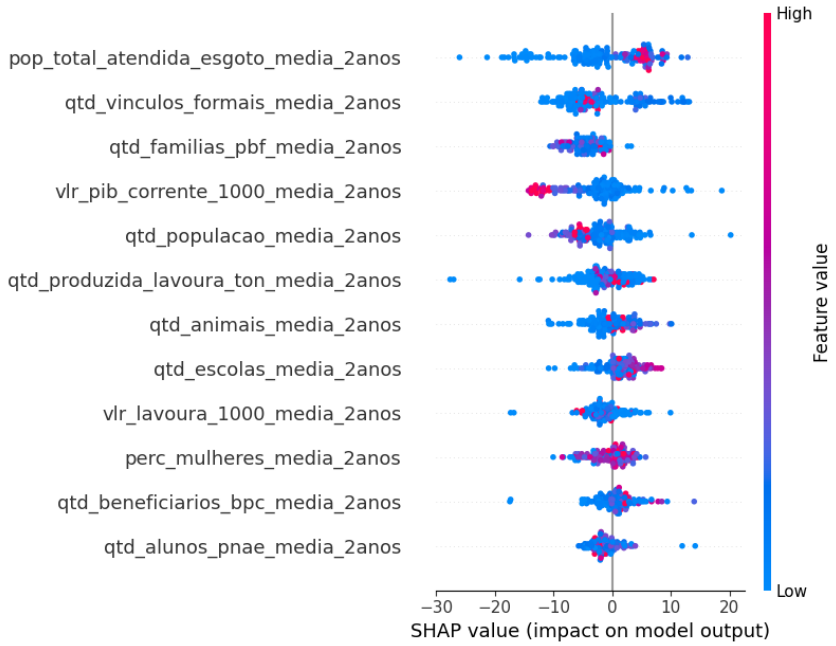
quantidade de alunos atendidos pelo PNAE, evidenciam a relevância de políticas sociais e de mecanismos de transferência de renda no contexto analisado. Como mostrado na Figura 5.6, a quantidade de pessoas atendidas pelo Programa Bolsa Família permaneceu relativamente estável no período selecionado, com exceção de 2021, sendo este um ano atípico devido à pandemia do Covid-19. É de importância reforçar a necessidade e importância das políticas de transferência de renda no enfrentamento de vulnerabilidades socioeconômicas.

Figura 6.17: SHAP para Norte (LGBM)



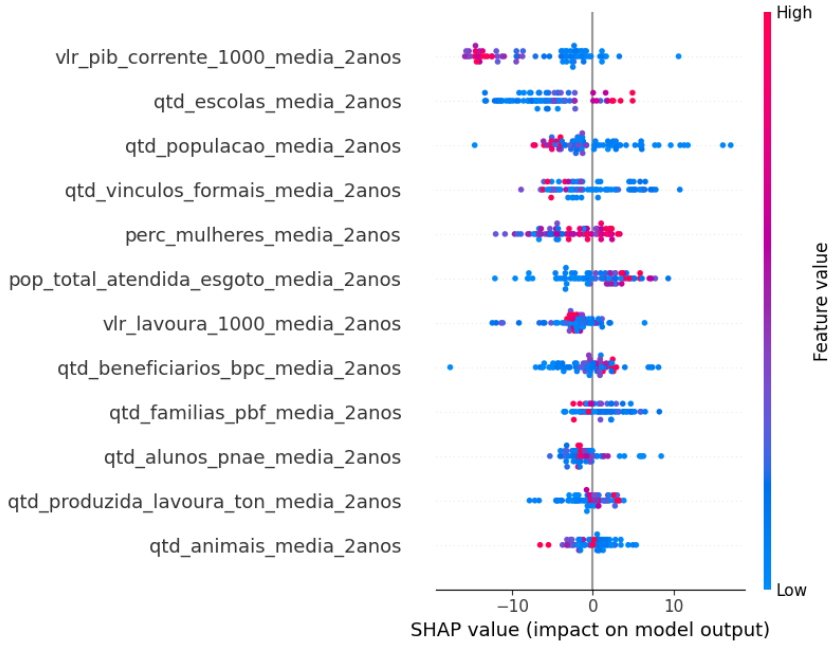
Fonte: De autoria própria.

Figura 6.18: SHAP para Nordeste (LGBM)



Fonte: De autoria própria.

Figura 6.19: SHAP para Centro-Oeste (LGBM)



Fonte: De autoria própria.

6.7 Modelos locais para as macrorregiões do país

Considerando os atributos médios dos dois anos anteriores a cada ano t, aplicou-se o algoritmo RFE e submeteu-se à busca por hiperparâmetros em conjuntos de dados

Tabela 6.9: Métricas dos modelos locais

Modelo	**LGBM**		***Random Forest***		***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	0,52	0,31	0,57	0,34	0,43	0,31
Sudeste	0,52	0,31	0,57	0,34	0,43	0,31
Sul	0,52	0,33	0,57	0,32	0,43	0,32

segregados por cada macrorregião do Brasil, obtendo-se, assim, modelos locais. A metodologia aplicada foi exatamente a já descrita neste trabalho. Os resultados das avaliações para as macrorregiões estão na Tabela 6.9. De modo geral, observa-se que, diferentemente do modelo global, o desempenho varia menos entre as 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. Neste, a melhor métrica foi MAPE de 0,54 no modelo LGBM, já no modelo local a melhor métrica MAPE foi de 0,35 no *Random Forest* 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 local o melhor valor é de 0,32, para *Random Forest* e *XGBoost*. Uma hipótese a ser levantada é que, conforme será mostrado na Seção 6.8, 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.

Ao se observar os atributos que foram selecionados para modelo em cada macrorregiões, vê-se que tanto a quantidade de atributos quanto os atributos selecionados variam entre modelos e entre macrorregiões, fortalecendo o argumento de que a relação entre os atributos explicativos e a taxa de internação por desnutrição não é homogênea no território nacional. Os dados dos atributos selecionados para cada modelo local para cada macrorregião são apresentados no Apêndice, Seção 8.

6.8 Redução de dimensionalidade

Com o objetivo de aprofundar a compreensão da influência da segmentação territorial na estrutura dos dados e de investigar possíveis padrões regionais não explicitamente capturados pelas métricas de desempenho dos modelos, aplicaram-se técnicas de redução de dimensionalidade aos atributos selecionados. Especificamente, foi utilizado PCA e t-SNE.

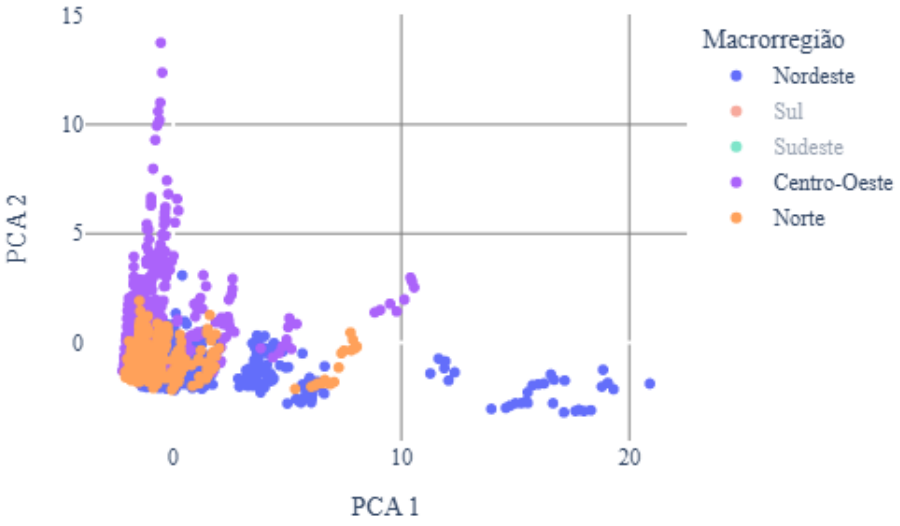
A Figura 6.20 apresenta a projeção PCA com dois componentes principais para as macrorregiões Norte, Nordeste e Centro-Oeste, enquanto a Figura 6.21 apresenta a mesma projeção restrita às macrorregiões Sul e Sudeste. Foi realizada esta separação para melhor visualização, pois há uma sobreposição em algumas partes do gráfico entre as diferentes macrorregiões, o que tornava difícil de visualizar as distinções. Além disso, como se observa no gráfico 6.21, as macrorregiões Sul e Sudeste apresentam características similares, de modo que sua distribuição é basicamente sobreposta uma a outra, de maneira bastante homogênea. Tal comportamento indica maior homogeneidade interna nessas macrorregiões, o que está alinhado com os resultados observados anteriormente, nos quais essas macrorregiões apresentaram melhor desempenho preditivo e menor variabilidade dos resíduos. Essa homogeneidade pode estar relacionada a níveis mais consistentes de renda, infraestrutura e cobertura de políticas públicas.

Observa-se que, para o conjunto excluindo Sul e Sudeste, não há uma separação tão clara entre as macrorregiões. Esse resultado sugere que, apesar das diferenças socioeconômicas conhecidas entre essas macrorregiões, os atributos selecionados apresentam padrões semelhantes quando analisados de forma conjunta e linear. As macrorregiões Norte e Nordeste, inclusive, possuem uma distribuição muito similar.

A Tabela 6.10 apresenta os atributos que mais contribuíram para a formação dos dois primeiros componentes principais. Observa-se que atributos como a população figura entre os principais contribuintes, assim como na Figura 6.12, porém, também se vê que outros atributos aparecem como relevantes na análise PCA, como a produção de lavoura, o que indica que ela é uma variável relevante para a segregação das macrorregiões, porém não necessariamente, nos modelos testados, tem bom poder explicativo sobre a desnutrição.

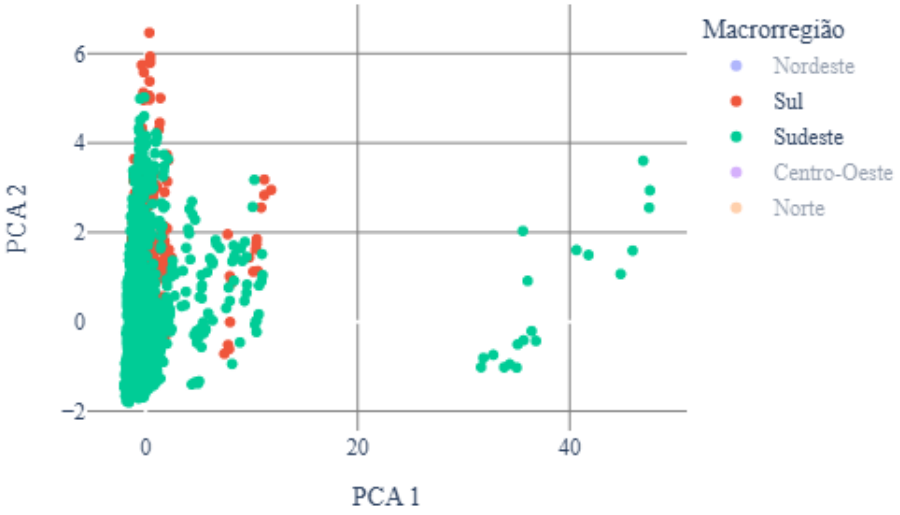
As Figuras 6.22 e 6.23 apresentam as projeções obtidas por meio do t-SNE para os mesmos subconjuntos regionais. Diferentemente do PCA, o t-SNE, por ser uma técnica não linear, evidencia agrupamentos locais mais definidos. Para as macrorregiões Norte,

Figura 6.20: PCA (exceto macrorregiões Sul e Sudeste)



Fonte: De autoria própria.

Figura 6.21: PCA (macrorregiões Sul e Sudeste)



Fonte: De autoria própria.

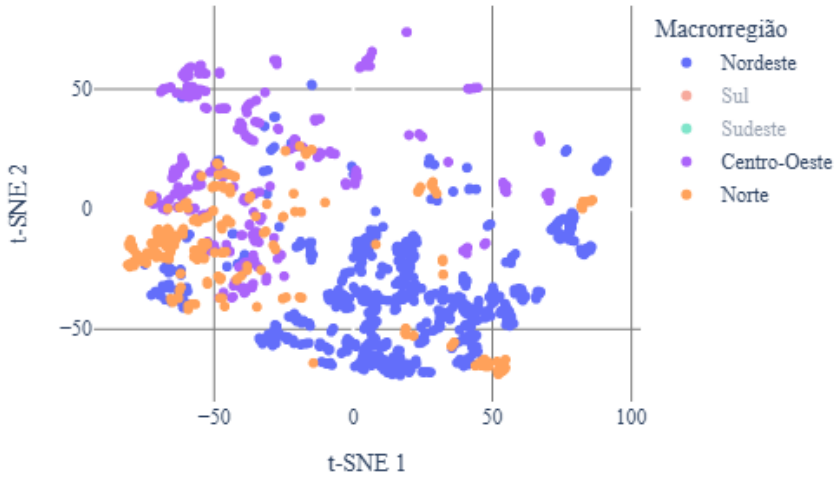
Tabela 6.10: Atributos que mais contribuem para PCA

Atributo	Valor
PCA1	
qtd_populacao_media_2anos	0,2810
qtd_alunos_ensino_basico_media_2anos	0,2809
qtd_alunos_pnae_media_2anos	0,2781
pop_total_atendida_agua_media_2anos	0,2773
pop_total_atendida_coleta_media_2anos	0,2772
PCA2	
vlr_lavoura_1000_media_2anos	0,4770
qtd_produzida_lavoura_ton_media_2anos	0,4343
vlr_media_salario_formal_media_2anos	0,4215
perc_adesao_politica_saneamento_media_2anos	0,4099
qtd_animais_media_2anos	0,3814

Nordeste e Centro-Oeste, observa-se a formação de pequenos agrupamentos dispersos, com maior heterogeneidade para a região Nordeste. No caso das macrorregiões Sul e Sudeste, o t-SNE novamente evidencia agrupamentos mais compactos e bem definidos, indicando maior similaridade entre as regiões do SUS pertencentes a essas macrorregiões.

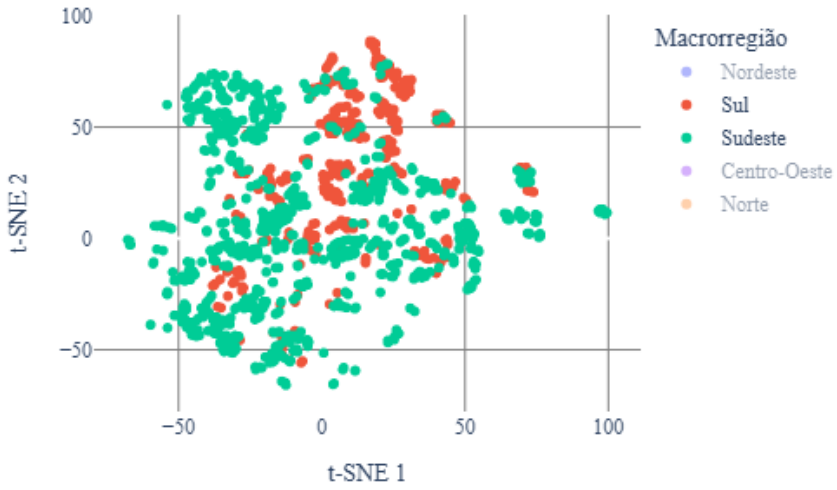
Em síntese, a análise de redução de dimensionalidade indica que a segmentação territorial exerce influência relevante na estrutura dos dados. Macrorregiões com maior homogeneidade socioeconômica e institucional, como Sul e Sudeste, tendem a formar agrupamentos mais consistentes, enquanto Norte, Nordeste e Centro-Oeste apresentam maior dispersão e heterogeneidade.

Figura 6.22: t-SNE (exceto macrorregiões Sul e Sudeste)



Fonte: De autoria própria.

Figura 6.23: t-SNE (macrorregiões Sul e Sudeste)



Fonte: De autoria própria.

7 CONCLUSÃO

Ao longo deste trabalho, foram desenvolvidos e avaliados modelos preditivos para a taxa de internação hospitalar anual por desnutrição no SUS, segmentada pelas regiões de saúde do Brasil, de acordo com o município de residência. Buscou-se investigar a viabilidade de modelar esse fenômeno a partir de variáveis socioeconômicas, demográficas, educacionais e de infraestrutura, além de compreender os fatores mais associados às internações por desnutrição.

Os experimentos realizados com *Random Forest*, *XGBoost* e LGBM mostraram que é possível obter capacidade preditiva para a taxa de internação por desnutrição. Embora, no conjunto de validação OOT, os modelos tenham apresentado métricas inferiores que no conjunto temporal de teste, 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%, indicando que os algoritmos conseguiram capturar padrões relevantes presentes nos dados históricos e nos atributos explicativos selecionadas.

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. Esse resultado reforça a importância de considerar o espaço geográfico como determinante social da saúde, bem como a necessidade de estratégias regionais diferenciadas no planejamento de políticas públicas voltadas à redução da insegurança alimentar e da desnutrição.

Os métodos de explicabilidade baseados em SHAP permitiram identificar os atributos com maior contribuição para as predições, destacando a relevância de variáveis relacionadas à renda, de programas de transferência de renda, de escolaridade e de infraestrutura básica. Uma das dificuldades encontradas na análise dos atributos foi a população atendida por esgoto tratado, que, a princípio, possui comportamento anti-intuitivo, talvez derivado do fato de a região Sul não possuir boas taxas de atendimento neste atributo. Assim, é um atributo com comportamento distinto dos outros, nos quais as melhores taxas de atributos socio-econômicos estão no Sul e Sudeste. O Norte, por exemplo, possui uma população total maior do que o Centro-Oeste, mas um atendimento de esgoto muito

menor, enquanto o Sudeste se destaca em termos de atendimento e população total.

Entre as limitações do estudo, destacam-se a dependência da qualidade e da periodicidade das bases públicas utilizadas, possíveis subnotificações nos registros de internação e a ausência de variáveis clínicas individuais, que poderiam refinar ainda mais as predições. Além disso, embora os modelos tenham apresentado desempenho satisfatório em termos globais, observa-se espaço para aprimoramento, especialmente na macrorregião Centro-Oeste, cujo desempenho dos modelos é muito abaixo das outras macrorregiões, o que levanta dúvidas, também, sobre os próprios dados governamentais para esta macrorregião. Em específico, seria de suma importância também analisar a facilidade de acesso à rede pública de saúde em cada local, 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.

Como perspectivas futuras, sugere-se a incorporação de novas fontes de dados, como indicadores de insegurança alimentar em nível domiciliar, dados de atenção primária e variáveis ambientais, além da experimentação com modelos intrinsecamente espaço-temporais, uma vez que esta é a natureza do problema.

Conclui-se, portanto, que o uso de modelos de aprendizado de máquina constitui uma abordagem promissora para a análise e predição da taxa de internação por desnutrição no SUS. Espera-se que os resultados e metodologias apresentados neste trabalho contribuam para o avanço de pesquisas na interface entre ciência de dados e saúde pública, bem como para o fortalecimento de estratégias de enfrentamento da desnutrição no Brasil.

8 APÊNDICE

Tabela 8.1: Atributos selecionados para cada modelo para o Centro-Oeste

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x		x
vlr_pbf_media_2anos	x		
qtd_familias_pbf_media_2anos	x		x
vlr_bpc_media_2anos	x		
qtd_beneficiarios_bpc_media_2anos	x	x	x
vlr_medio_salario_formal_media_2anos	x		
qtd_vinculos_formais_media_2anos	x		x
qtd_supermercados_media_2anos	x		
perc_mulheres_media_2anos	x	x	
qtd_populacao_media_2anos	x	x	x
qtd_escolas_media_2anos	x		x
qtd_alunos_ensino_basico_media_2anos	x		
qtd_alunos_pnae_media_2anos			
qtd_animais_media_2anos	x		
vlr_lavoura_1000_media_2anos	x		x
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos	x		
pop_total_atendida_esgoto_media_2anos	x		
pop_total_atendida_coleta_media_2anos	x		
perc_adesao_politica_saneamento_media_2anos	x		

Tabela 8.2: Atributos selecionados para cada modelo para o Nordeste

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x	x	x
vlr_pbf_media_2anos	x		x
qtd_familias_pbf_media_2anos	x		x
vlr_bpc_media_2anos	x		x
qtd_beneficiarios_bpc_media_2anos	x	x	x
vlr_medio_salario_formal_media_2anos	x	x	x
qtd_vinculos_formais_media_2anos	x	x	x
qtd_supermercados_media_2anos	x		x
perc_mulheres_media_2anos	x	x	x
qtd_populacao_media_2anos	x	x	x
qtd_escolas_media_2anos	x	x	x
qtd_alunos_ensino_basico_media_2anos	x	x	x
qtd_alunos_pnae_media_2anos			x
qtd_animais_media_2anos	x	x	x
vlr_lavoura_1000_media_2anos	x	x	x
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos	x		x
pop_total_atendida_esgoto_media_2anos	x	x	x
pop_total_atendida_coleta_media_2anos	x	x	x
perc_adesao_politica_saneamento_media_2anos	x		x

Tabela 8.3: Atributos selecionados para cada modelo para o Norte

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x	x	x
vlr_pbf_media_2anos	x		x
qtd_familias_pbf_media_2anos	x		x
vlr_bpc_media_2anos	x		x
qtd_beneficiarios_bpc_media_2anos	x	x	x
vlr_medio_salario_formal_media_2anos	x	x	x
qtd_vinculos_formais_media_2anos	x	x	x
qtd_supermercados_media_2anos	x		x
perc_mulheres_media_2anos	x	x	x
qtd_populacao_media_2anos	x	x	x
qtd_escolas_media_2anos	x	x	x
qtd_alunos_ensino_basico_media_2anos	x	x	x
qtd_alunos_pnae_media_2anos			x
qtd_animais_media_2anos	x	x	x
vlr_lavoura_1000_media_2anos	x	x	x
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos	x		x
pop_total_atendida_esgoto_media_2anos	x	x	x
pop_total_atendida_coleta_media_2anos	x	x	x
perc_adesao_politica_saneamento_media_2anos	x		x

Tabela 8.4: Atributos selecionados para cada modelo para o Sudeste

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x	x	x
vlr_pbf_media_2anos	x		x
qtd_familias_pbf_media_2anos	x		x
vlr_bpc_media_2anos	x		x
qtd_beneficiarios_bpc_media_2anos	x	x	x
vlr_medio_salario_formal_media_2anos	x	x	x
qtd_vinculos_formais_media_2anos	x	x	x
qtd_supermercados_media_2anos	x		x
perc_mulheres_media_2anos	x	x	x
qtd_populacao_media_2anos	x	x	x
qtd_escolas_media_2anos	x	x	x
qtd_alunos_ensino_basico_media_2anos	x	x	x
qtd_alunos_pnae_media_2anos			x
qtd_animais_media_2anos	x	x	x
vlr_lavoura_1000_media_2anos	x	x	x
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos	x		x
pop_total_atendida_esgoto_media_2anos	x	x	x
pop_total_atendida_coleta_media_2anos	x	x	x
perc_adesao_politica_saneamento_media_2anos	x		x

Tabela 8.5: Atributos selecionados para cada modelo para o Sul

Atributo	**LGBM**	***Random Forest***	***XGBoost***
vlr_pib_corrente_1000_media_2anos	x	x	x
vlr_pbf_media_2anos			
qtd_familias_pbf_media_2anos	x		x
vlr_bpc_media_2anos	x		
qtd_beneficiarios_bpc_media_2anos		x	x
vlr_medio_salario_formal_media_2anos	x	x	x
qtd_vinculos_formais_media_2anos	x	x	x
qtd_supermercados_media_2anos			
perc_mulheres_media_2anos	x	x	x
qtd_populacao_media_2anos		x	x
qtd_escolas_media_2anos	x	x	x
qtd_alunos_ensino_basico_media_2anos		x	x
qtd_alunos_pnae_media_2anos			x
qtd_animais_media_2anos	x	x	
vlr_lavoura_1000_media_2anos			
qtd_produzida_lavoura_ton_media_2anos	x	x	x
pop_total_atendida_agua_media_2anos	x		
pop_total_atendida_esgoto_media_2anos	x	x	x
pop_total_atendida_coleta_media_2anos			
perc_adesao_politica_saneamento_media_2anos			

REFERÊNCIAS

ALMALKI, A. et al. Regression analysis for covid-19 infections and deaths based on food access and health issues. **Healthcare**, v. 10, n. 2, 2022. ISSN 2227-9032. Available from Internet: <https://www.mdpi.com/2227-9032/10/2/324>.

ALMALKI, A. et al. Geospatial and machine learning regression techniques for analyzing food access impact on health issues in sustainable communities. **ISPRS International Journal of Geo-Information**, v. 10, n. 11, 2021. ISSN 2220-9964. Available from Internet: <https://www.mdpi.com/2220-9964/10/11/745>.

ALMEIDA, G. et al. Hospital length-of-stay prediction using machine learning algorithms—a literature review. **Applied Sciences**, v. 14, n. 22, 2024. ISSN 2076-3417. Available from Internet: <https://www.mdpi.com/2076-3417/14/22/10523>.

AWAN, A. A. **An Introduction to SHAP Values and Machine Learning Interpretability**. DataCamp, 2023. Accessed: 2025-11-24. Available from Internet: <https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability>.

AWAN, A. A. **Introdução ao t-SNE: Redução de dimensionalidade não linear e visualização de dados**. 2024. Acessado em: 22 dez. 2025. Available from Internet: <https://www.datacamp.com/pt/tutorial/introduction-t-sne>.

BELCIC, I.; STRYKER, C. **What is feature selection?** 2025. Available from Internet: <https://www.ibm.com/think/topics/feature-selection>.

BERRAR, D. Cross-validation. In: ____. [S.l.: s.n.], 2018. ISBN 9780128096338.

CERQUEIRA, V.; TORGO, L.; MOZETIč, I. Evaluating time series forecasting models: an empirical study on performance estimation methods. **Machine Learning**, Springer Science and Business Media LLC, v. 109, n. 11, p. 1997–2028, oct. 2020. ISSN 1573-0565. Available from Internet: <http://dx.doi.org/10.1007/s10994-020-05910-7>.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM, 2016. (KDD '16), p. 785–794. Available from Internet: <http://dx.doi.org/10.1145/2939672.2939785>.

CHEN, X.-w.; JEONG, J. C. Enhanced recursive feature elimination. In: **Sixth International Conference on Machine Learning and Applications (ICMLA 2007)**. [S.l.: s.n.], 2007. p. 429–435.

CZAJKOWSKI, M.; KRETOWSKI, M. The role of decision tree representation in regression problems – an evolutionary perspective. **Applied Soft Computing**, v. 48, p. 458–475, 2016. ISSN 1568-4946. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1568494616303325>.

FAO et al. **The State of Food Security and Nutrition in the World 2025 – Addressing High Food Price Inflation for Food Security and Nutrition**. Rome, 2025. Available from Internet: <https://openknowledge.fao.org/server/api/core/bitstreams/4eed749b-81f8-49c9-ba32-f09c66988d54/content/cd6008en.html#gsc.tab=0>.

FRAZIER, P. I. **A Tutorial on Bayesian Optimization**. 2018. Available from Internet: <https://arxiv.org/abs/1807.02811>.

HIBBELN, M. T. et al. Model validation for pooled cross-sectional data: Out-of-sample vs. out-of-time. **SSRN Electronic Journal**, April 29 2025. Available at SSRN. Available from Internet: <https://ssrn.com/abstract=5241694>.

HUA, S. et al. Barriers and enablers for equitable healthy food access in baltimore carryout restaurants: A qualitative study in healthy food priority areas. **Nutrients**, v. 16, n. 17, p. 3028, 2024.

IBM. **O que é a análise de componentes principais (PCA)?** 2024. Acessado em: 22 dez. 2025. Available from Internet: <https://www.ibm.com/br-pt/think/topics/principal-component-analysis>.

JIA, W. et al. Feature dimensionality reduction: a review. **Complex & Intelligent Systems**, v. 8, n. 3, p. 2663–2693, 2022. ISSN 2198-6053. Available from Internet: <https://doi.org/10.1007/s40747-021-00637-x>.

K, S.; SARATHAMBEKAI, S. Machine learning based prediction model for health care sector - a survey. In: **2021 Innovations in Power and Advanced Computing Technologies (i-PACT)**. [S.l.: s.n.], 2021. p. 1–7.

KHARKAR, R. Decision tree regressor explained: A visual guide with code examples. **Towards Data Science**, July 2020. Acesso: 2025-11-07. Available from Internet: <https://towardsdatascience.com/decision-tree-regressor-explained-a-visual-guide-with-code-examples-fbd2836c3bef>.

KUMAR, S.; SRIVASTAVA, M.; PRAKASH, V. Advanced hybrid prediction model: Optimizing lightgbm, xgboost, lasso regression, and random forest with bayesian optimization. **Journal of Theoretical and Applied Information Technology**, v. 102, n. 9, 2024. ISSN 1817-3195. Available from Internet: <https://jatit.org/volumes/Vol102No9/32Vol102No9.pdf>.

LI, J. Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? then what? **PLOS ONE**, Public Library of Science, v. 12, n. 8, p. e0183250, 2017. Available from Internet: <https://doi.org/10.1371/journal.pone.0183250>.

LI, J. et al. Feature selection: A data perspective. **ACM Comput. Surv.**, Association for Computing Machinery, New York, NY, USA, v. 50, n. 6, dec. 2017. ISSN 0360-0300. Available from Internet: <https://doi.org/10.1145/3136625>.

LIU, H.; MOTODA, H. **Feature Selection for Knowledge Discovery and Data Mining**. 1. ed. Boston: Springer, 2012.

MAULUD, D.; ABDULAZEEZ, A. M. A review on linear regression comprehensive in machine learning. **Journal of Applied Science and Technology Trends**, v. 1, n. 2, p. 140–147, Dec. 2020. Available from Internet: <https://jastt.org/index.php/jasttpath/article/view/57>.

NICHOLS, J. A.; CHAN, H. W. H.; BAKER, M. A. B. Machine learning: applications of artificial intelligence to imaging and diagnosis. **Biophys Rev**, Germany, v. 11, n. 1, p. 111–118, sep. 2018.

PLEVRIS, V. et al. Investigation of performance metrics in regression analysis and machine learning-based prediction models. **Scipedia**, Nordic Association of Computational Mechanics, v. 50, n. 6, 2022. ISSN 0360-0300. Available from Internet: <https://doi.org/10.23967/eccomas.2022.155>.

RAO, B. et al. Machine learning in predicting child malnutrition: A meta-analysis of demographic and health surveys data. **International Journal of Environmental Research and Public Health**, v. 22, n. 3, 2025. ISSN 1660-4601. Available from Internet: <https://www.mdpi.com/1660-4601/22/3/449>.

SALIH, A. et al. A perspective on explainable artificial intelligence methods: Shap and lime. **Advanced Intelligent Systems**, v. 7, June 27 2024. Available at SSRN. Available from Internet: <https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.202400304>.

SILVA, T. L. C. da et al. 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**, v. 31, n. 1, p. 203–218, Mar. 2025. Available from Internet: <https://journals-sol.sbc.org.br/index.php/jbcs/article/view/3100>.

SIMEON, Y. **Alimentação escolar é a principal refeição para 56% dos estudantes do Grande Rio, revela pesquisa**. 2023. Acessado em: 31 dez. 2025. Available from Internet: <https://alimentacaoescolar.org.br/noticias/2023/03/17/alimentacao-escolar-e-a-principal-refeicao-para-56-dos-estudantes-do-grande-rio-revela-pesquisa/>.

VALADARES, A.; SOUZA, M.; ESTEVES, M. **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)**. [S.l.], 2023. Available from Internet: <https://www.gov.br/mds/pt-br/caisan/monitoramento-da-san/MapaInSAN_20172022.pdf>.

WISSENBACH, T.; BICEV, J.; VANETTI, V. **Relatório do Observatório Brasileiro das Desigualdades**. [S.l.], 2024. Available from Internet: <https://combateasdesigualdades.org/wp-content/uploads/2024/09/RELATORIO_2024_v3-1.pdf>.

WISTUBA, M.; SCHILLING, N.; SCHMIDT-THIEME, L. Learning hyperparameter optimization initializations. In: **2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.: s.n.], 2015. p. 1–10.