Dengue e Desigualdade: Avaliação da Incidência em Classes Sociais no Brasil

Pedro Augusto Gomes Minaré[1]; Ricardo Limongi França Coelho[2]

[1] Desenvolvedor Backend, Goiânia, Goiás, Brasil.
[2] MBA USP Esalq. Orientador Ricardo Limongi França Coelho, Universidade Federal de Goiás, Faculdade de Administração, Ciências Contábeis e Economia. Universidade Federal de Goiás - UFG - Campus Samambaia - Campus Universitário - 74690900 - Goiânia, GO - Brasil - Telefone: (62) 35211590
*autor correspondente: pedrominare94@gmail.com

Dengue e Desigualdade: Avaliação da Incidência em Classes Sociais no Brasil

Resumo

A dengue é considerada um grande problema de saúde pública no Brasil, impulsionada pelas condições climáticas tropicais e população 87% urbana (IBGE, 2022) que impactam na densidade populacional dos municípios e fatores sociais. A relação entre vulnerabilidade social e contaminação por dengue torna-se evidente quando populações em situação de maior precariedade, com acesso a saneamento básico limitado, bem como infraestrutura urbana e serviços de saúde precários, estão mais expostas ao vetor da doença. Assim, variáveis socioeconômicas como condições de moradia, renda e acesso a saneamento influenciam a incidência de dengue em cidades como o Rio de Janeiro, RJ, exigindo políticas públicas eficazes. O objetivo do trabalho é investigar se, e quais indicadores socioeconômicos estão associados a uma maior probabilidade de contaminação por dengue nos municípios brasileiros, a fim de estimar quais classes sociais apresentam maior vulnerabilidade para a contaminação. O desempenho obtido nas predições do modelo Random Forest demonstraram que as variáveis como renda média, acesso à infraestrutura de saneamento básica, índice de escolaridade e condições habitacionais estão associadas ao risco de contaminação por dengue. A maioria dos municípios possuem probabilidade moderada de contaminação, variando entre 20% e 60%, enquanto alguns municípios, especialmente nas regiões Nordeste, Norte e Centro-Oeste, obtiveram probabilidades superiores a 70%. Municípios com piores indicadores socioeconômicos apresentaram maior probabilidade de incidência da doença, o que sugere que a vulnerabilidade social é um fator crítico no contexto epidemiológico da dengue. Portanto, a utilização de modelos preditivos como o Random Forest se mostrou útil para identificar regiões prioritárias para ações de saúde pública, reafirmando a importância de se integrar métodos de análise de dados e aprendizado de máquina na elaboração de estratégias de prevenção e controle de arboviroses.

Palavras-chave: dengue; incidência; socioeconômica; probabilidade; contaminação, Random Forest;

Introdução

O Brasil é o país que contém o maior número de casos de dengue no mundo, com 6296795 casos de suspeita e 3040736 casos confirmados (OMS, 2024). De acordo com Causa et al (2020) as condições existentes em regiões de vulnerabilidade social influenciam nos riscos de surtos de doenças infecciosas e causadas por vetores, como as arboviroses, tendo sua presença associada a características ambientais. Diante disso, sendo a dengue um dos grandes problemas de saúde pública do Brasil (Ministério da Saúde), onde as condições climáticas tropicais e os altos índices de crescimento urbano impulsionam os riscos de surtos, este estudo traz a proposta de investigar se as condições socioeconômicas são um fator influente na incidência de dengue em municípios brasileiros.

Conforme Sobral de Almeida et al. (2009), ao relacionar a ocorrência de dengue com variáveis socioeconômicas foi possível concluir que existe uma relação entre a contaminação de dengue de um indivíduo e fatores como condições de moradia, renda e principalmente o acesso a saneamento básico. O artigo identificou como variável mais significativamente associada à dengue o percentual de domicílios ligados à rede sanitária geral.

Com a análise de dados de contaminação por município é possível desenvolver modelos que estimem a probabilidade de contaminação por dengue em diferentes regiões, permitindo que tanto o direcionamento de recursos para áreas de risco quanto políticas públicas de combate à proliferação da doença sejam criadas de forma mais eficaz.

Indicadores de saneamento e outras variáveis socioeconômicas relacionadas à vulnerabilidade social mostraram eficácia para classificar a chance de contaminação por município, considerando sua população total.

Foi utilizada uma abordagem quantitativa por meio da modelagem estatística, utilizando o modelo linear generalizado (GLM) com distribuição binomial, tendo como variável dependente a ocorrência de contaminação (CONTAMINACAO_INDICE) e como preditoras os indicadores socioeconômicos disponíveis no dataset.

Metodologia

Para a análise observacional e estatística, os dados foram coletados a partir de observações, como o número de casos de dengue por município, municípios de todos os estados brasileiros com exceção do Distrito Federal e indicadores socioeconômicos dos municípios, utilizados para identificar associações estatísticas, como medir a força e direção das relações entre as variáveis, como ilustra Hosmer et al (2013), Lemeshow e Sturdivant, ao enfatizarem que a regressão logística é particularmente útil para dados provenientes de estudos observacionais, como pesquisas epidemiológicas, a fim de medir associações entre variáveis.

Os dados foram coletados de fontes oficiais e públicas, correspondentes especificamente ao ano de 2010. Os índices de incidência de dengue foram obtidos inicialmente por meio de relatórios da Secretaria de Estado da Saúde (SES), especificamente para o estado de Goiás, e os indicadores socioeconômicos, como renda média familiar, cobertura de saneamento e densidade populacional foram extraídos de bases disponibilizadas pelo IBGE (Instituto Brasileiro de Geografia e Estatística) para as unidades federativas.

A escolha do estado de Goiás como objeto inicial de análise se justifica por sua localização geográfica estratégica no Centro-Oeste brasileiro, caracterizada pelo clima tropical, alternância de períodos de seca e chuvas intensas e condições ambientais altamente propícias à proliferação do vetor da doença. Ainda, Goiás apresenta altos índices de casos notificados de dengue em comparação com outros estados, o que torna a região relevante para investigar a relação entre vulnerabilidade social e incidência da doença em contextos urbanos e semiurbanos.

Nem todos os estados brasileiros disponibilizam os dados de contaminação de dengue em um sistema como o SES. Assim, para os demais estados, os dados foram coletados do Datasus, no sistema Tabnet.

A Tabela 1 apresenta os indicadores socioeconômicos a serem utilizados para análise:

Tabela de Variáveis Socioeconômicas	
Variável	**Descrição**
População residente em casa própria, saneamento inadequado e seu rendimento mensal total per capita	Grupo de variáveis quantitativas que representam a população residente em domicílios particulares permanentes, com saneamento inadequado, e a proporção de pessoas por classes selecionadas de rendimento mensal total domiciliar per capita nominal, mediante os municípios e as classes de tamanho da população dos municípios.
População Total do Município	Variável quantitativa representando a população total de cada município.
População residente em casa própria e rendimento mensal per capita	Grupo de variáveis quantitativas que representam a população residente em domicílios particulares permanentes, e a proporção de pessoas residentes em domicílios particulares permanentes, por situação do domicílio, e classes selecionadas de rendimento mensal total domiciliar per capita nominal, mediante os municípios e as classes de tamanho da população dos municípios.
Proporção do rendimento mensal da população residente em casa própria com mais de 10 anos, por cor ou raça, de acordo com proporções do salário mínimo.	Grupo de variáveis quantitativas que representam a razão entre as médias do rendimento mensal total nominal, das pessoas de no mínimo 10 anos de idade residentes em domicílios particulares permanentes, por cor ou raça, mediante os municípios e por proporções do salário mínimo.

Média e mediana do rendimento mensal da população com mais de 10 anos, residente em casa própria, por sexo.	Grupo de variáveis quantitativas que representam o valor médio e mediano do rendimento mensal total nominal das pessoas com no mínimo 10 anos de idade, residentes em domicílios particulares permanentes, por sexo, e a razão entre valor médio e mediano do rendimento mensal total nominal de homens e mulheres, mediante os municípios e classes de tamanho da população dos municípios.
Média e quartis do rendimento mensal per capita por situação do domicílio.	Grupo de variáveis quantitativas que representam o valor médio e quartis do rendimento mensal total domiciliar per capita nominal, em virtude da por situação do domicílio, mediante os municípios e as classes de tamanho da população dos municípios.
UF	Unidade federativa do município de residência do indivíduo.
Município	Variável qualitativa que descreve o município de residência do indivíduo.
Ano epidemiológico	Variável quantitativa que representa o ano em que foi notificada a contaminação.
Semana Epidemiológica	Variável quantitativa que representa o padrão da OMS para organizar dados temporais de contaminação de dengue.
Data dos Sintomas	Variável quantitativa que representa a data em que os sintomas iniciaram.
Data da Notificação	Variável quantitativa que representa a data da notificação da contaminação ou suspeita de contaminação de dengue.
Classificação	Variável qualitativa que representa a classificação quanto ao estado de saúde do indivíduo, entre DENGUE, DENGUE COM SINAIS DE ALARME, DENGUE GRAVE,

	DESCARTADO, INCONCLUSIVO e SUSPEITO.
Critério de confirmação	Variável qualitativa que representa o critério utilizado para a confirmação da contaminação de dengue, entre CLÍNICO-EPIDEMIOLÓGICO, EM INVESTIGAÇÃO, LABORATÓRIO e SUSPEITO.
Evolução	Variável qualitativa que representa o estado de saúde do indivíduo no término do tratamento, entre CURA, IGNORADO, ÓBITO EM INVESTIGAÇÃO, ÓBITO PELO AGRAVO, ÓBITO POR OUTRAS CAUSAS e SUSPEITO.

Tabela 1: Tabela de Variáveis Socioeconômicas do site do IBGE, Censo de 2010, e variáveis de contaminação de dengue do site do Governo de Goiás, por município.

A fim de organizar a visibilidade das variáveis na matriz de correlação, índices foram atribuídos às variáveis. O índice F1 representa a população total de cada município, enquanto os índices de F2 a F34 representam os indicadores socioeconômicos. O índice F33 representa a variável quantitativa dos casos de dengue por município, e o índice F34 representa a variável dicotômica CONTAMINACAO_INDICE.

Os dados que correspondem aos índices de contaminação de dengue consistem em um arquivo csv contendo observações com base no município de residência do indivíduo, ano e semana epidemiológica de contaminação, data dos sintomas, data da notificação, classificação, critério de confirmação e evolução do caso, especificamente para o estado de Goiás.

A técnica utilizada para estabelecer uma relação entre cada variável socioeconômica e a incidência de dengue será a Regressão Logística Binária que, de acordo com Hosmer et al (2013), Lemeshow e Sturdivant, tem como objetivo modelar a relação entre variáveis independentes, como os indicadores de vulnerabilidade social, e a ocorrência de contaminação de dengue, variável dependente binária, a fim de estabelecer a probabilidade de ocorrência de contaminação de dengue e da não-contaminação.

A Regressão Logística Binária é um modelo que descreve a probabilidade da ocorrência de um evento e um não-evento, ou seja, de uma observação pertencer a uma de duas classes possíveis com base em uma ou mais variáveis independentes (Hosmer et al, 2013). A função logística é utilizada para mapear o resultado de uma combinação linear de variáveis independentes, de valor 0 ou 1, representando a probabilidade do resultado ser 1.

$$p_i = \frac{1}{1+e^{-Z_i}} = \frac{1}{1+e^{-(\propto+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_k X_{ki})}}$$

Onde:

- p_i é a probabilidade de ocorrência do evento (variável dependente ser 1).
- $-Z_i$ (logito) é o logaritmo natural da chance de ocorrência de 1.
- $\propto$ é o intercepto, definido como a saída do modelo quando todos os preditores são iguais a zero.
- βi são os coeficientes das variáveis independentes X representando seu impacto no modelo.

Visto que o modelo exige que haja uma variável dependente binária, foi necessário determinar uma razão entre a quantidade de indivíduos contaminados com dengue, a média ponderada entre os índices socioeconômicos dos municípios e a população total do município, de modo que o parâmetro binário fosse estabelecido.

A variável CONTAMINACAO_INDICE foi atribuída ao dataset como sendo a razão entre a quantidade de indivíduos contaminados em um determinado município, a população total e a média ponderada de todos os índices socioeconômicos. Se a razão resultar em um valor maior ou igual a 1, a variável será atribuída como 1. Caso resulte em um valor menor que 1, a variável será atribuída como 0.

$$R = \frac{DENGUE \times \left(\frac{\sum_{i=1}^{n} w_i I_i}{\sum_{i=1}^{n} w_i} \right)}{POPULAÇÃO\ TOTAL}$$

Cada índice socioeconômico I_i é ponderado por um peso w_i que reflete a sua importância relativa. Os pesos de cada variável socioeconômica foram determinados com base nos resultados apresentados no artigo de Sobral de Almeida et al. (2009). O artigo concluiu que as variáveis que fazem referência a problemas relacionados a domicílios ligados à rede sanitária geral tendem a influenciar decisivamente no aumento do risco de contaminação de dengue. Assim, foram atribuídos pesos relativamente mais significantes nas variáveis que representam proporções de domicílios ligados à rede sanitária e menos significantes às demais variáveis.

Embora os pesos não tenham sido derivados por técnicas estatísticas formais (como PCA ou regressão regularizada), eles foram definidos com base em evidências consolidadas na literatura, garantindo coerência com o arcabouço teórico da epidemiologia crítica e da determinação social da saúde. O objetivo foi construir uma métrica sintética que priorizasse a interpretação substantiva dos indicadores, permitindo avaliar não apenas a carga da doença, mas também sua associação com desigualdades estruturais.

O Log-Likelihood (LL), ou função de verossimilhança, oriunda da função de probabilidade da distribuição de Bernoulli, é o indicador de eficiência do modelo que define o quanto o modelo deve acertar as previsões de índices de contaminação, ou seja, o quanto o modelo é aderente aos índices socioeconômicos quanto à chance de um indivíduo ser contaminado com dengue ou não em um dado município.

$$LL = \sum_{i=1}^{n} \left\{ \left[\left(Y_i\right). \ln ln \left(\frac{1}{1+e^{-Z_i}} \right) \right] + \left[\left(1 - Y_i\right). ln \left(\frac{1}{1+e^{Z_i}} \right) \right] \right\}$$

Quanto mais próximo de zero, melhor é a eficiência do modelo para prever os acertos em um dado conjunto de dados no qual o modelo utiliza.

Com a variável dependente binária foi possível observar que a taxa de contaminação corresponde a 72,36% das observações do dataset como sendo indivíduos não contaminados e 27,64% de indivíduos contaminados para o estado de Goiás. Antes da elaboração do modelo preditivo, foi feita uma análise exploratória dos dados disponíveis. Nesta etapa, destacaram-se as variáveis altamente correlacionadas com a variável dependente CONTAMINACAO_INDICE, em especial as variáveis TOTAL e DENGUE que, ainda que informativas, poderiam causar o problema conhecido como vazamento de informação. Isso ocorre quando informações que estariam disponíveis apenas após o evento de interesse são utilizadas para treinar o modelo, gerando uma estimativa de desempenho artificialmente otimista. Para evitar esse viés, as variáveis citadas foram removidas do conjunto de dados utilizado para treinamento. Ainda, variáveis categóricas como UF e MUNICIPIO também foram excluídas do treinamento, porque representavam identificadores geográficos e não contribuíram diretamente para a predição do risco de contaminação.

O objetivo do trabalho é apontar os indicadores socioeconômicos de municípios brasileiros que mais influenciam a probabilidade de indivíduos serem contaminados com dengue a fim de estabelecer métricas que descrevem quais classes sociais são mais vulneráveis à contaminação. Nesse sentido foi escolhido o algoritmo Random Forest, uma técnica de aprendizado supervisionado baseada em ensemble de árvores de decisão, possibilitando resultados mais robustos para um dataset com mais observações.

A opção de se utilizar o Random Forest se justifica por algumas razões, como o fato de que o conjunto de dados utilizado contém uma grande quantidade de variáveis socioeconômicas, muitas das quais podendo ser fortemente correlacionadas ou apresentar relações complexas e não lineares com o risco de contaminação de dengue. O Random Forest demonstra eficácia nesse cenário, sendo capaz de capturar interações não lineares entre variáveis sem necessidade de especificação explícita dessas relações (Breiman, 2001). Ainda, o modelo se apresenta robusto a problemas comuns em dados

socioeconômicos, como a presença de outliers e variáveis pouco informativas, em virtude do seu mecanismo de amostragem aleatória de atributos durante a construção de cada árvore.

Outro fator é a dimensão do conjunto de dados. Visto que mais de 5500 municípios foram analisados, o Random Forest foi capaz de oferecer excelente escalabilidade e capacidade de generalização, ao minimizar riscos de overfitting através da agregação de múltiplos classificadores. A capacidade intrínseca de calcular a importância relativa de cada variável faz do Random Forest um modelo que permitiu a identificação dos indicadores socioeconômicos mais relevantes para a previsão do risco de contaminação de dengue, fornecendo informações valiosas para o campo da saúde pública e para o planejamento de políticas públicas.

O processo de construção do modelo foi estruturado em etapas essenciais para obter a melhor performance e validade dos resultados. Inicialmente, o conjunto de dados foi dividido em subconjuntos de treinamento e teste utilizando a função train_test_split da biblioteca scikit-learn. Nesta divisão, atribuiu-se 70% dos dados para o treinamento e 30% para o teste (test_size = 0.3), certificando-se de que a amostra de teste fosse representativa da distribuição original da variável resposta por meio do parâmetro stratify=y.

A estratificação é recomendada para problemas que envolvem classificação, pois garante a preservação da proporção das classes em ambos os conjuntos (Géron, 2019). Para a reprodutibilidade dos resultados, foi fixado o parâmetro random_state=42, de forma a permitir que demais execuções do código resultem na mesma partição dos dados.

O Random Forest é um algoritmo de aprendizado supervisionado com base no conceito de "bagging" (Breiman, 2001), combinando diversos classificadores de árvores de decisão, agregando seus resultados para melhorar a precisão e reduzir o risco de overfitting. A fim de melhorar o desempenho do modelo, foi realizada a otimização dos seus hiperparâmetros por meio da técnica de busca em grade GridSearchCV. A busca em grade consiste na avaliação exaustiva de todas as combinações possíveis dos hiperparâmetros especificados em uma malha predefinida.

Os hiperparâmetros explorados foram o número de árvores na floresta (n_estimators), testando valores 100 e 200, indicando robustez no modelo mesmo que aumentando o custo computacional; max_depth, indicando a profundidade máxima de cada árvore como sendo 5, 10 e que as árvores cresçam até que todas as folhas sejam puras ou contenham menos amostras que o mínimo necessário, limitando o overfitting; min_samples_split, representando o número mínimo de amostras exigido para dividir um nó interno, de valores 2 e 5; min_samples_leaf, indicando o número mínimo de amostras exigido um nó folha, de valores 1 e 2, atuando como regularizador, evitando que a árvore se torne mais simples.

A avaliação das combinações de hiperparâmetros foi realizada pela técnica de validação cruzada com 5 dobras (cv = 5), garantindo que o modelo fosse treinado e validado em diferentes subconjuntos dos dados, diminuindo a variabilidade da avaliação.

A métrica utilizada para avaliar o desempenho dos modelos foi a ROC AUC (scoring = 'roc_auc'), medindo a capacidade do classificador de distinguir entre as classes. A escolha da métrica ROC AUC é adequada para problemas de classificação desbalanceada, pois considera a taxa de verdadeiros positivos e a taxa de falsos positivos em todas as possíveis configurações de limiar (Fawcett, 2006).

O parâmetro n_jobs = -1 foi utilizado para viabilizar a execução paralela em todos os processadores disponíveis, acelerando o processo de busca. O parâmetro verbose = 1 foi ativado para fornecer feedback detalhado sobre o progresso da busca durante sua execução.

Resultados e Discussão

A pesquisa teve como base um dataset contendo as variáveis independentes como sendo as variáveis socioeconômicas citadas, o total da população de cada município e o total de indivíduos contaminados por município, com proporções e valores absolutos, e a variável dependente sendo o índice CONTAMINACAO_INDICE. Inicialmente a análise foi realizada somente para os municípios do estado de Goiás, seus índices socioeconômicos e a quantidade de indivíduos contaminados por município. Posteriormente a mesma análise foi aplicada para os municípios com dados de contaminação de todas as unidades federativas exceto o distrito federal.

A análise exploratória das variáveis demonstrou um cenário de desigualdade que expõe grande parte da população à vulnerabilidade no que se refere à contaminação por dengue. A variável que se refere à proporção de famílias residentes em domicílios com saneamento semi-adequado demonstrou uma distribuição razoavelmente uniforme, onde muitos municípios concentram proporções de valores entre 20% e 60%. A variável referente à proporção de indivíduos residentes em áreas com saneamento inadequado demonstrou que embora a maioria dos municípios tenha a proporção de domicílios com condições sanitárias inadequadas baixa, com concentração entre 0% e 10%, existem municípios com índices superiores a 40%. O saneamento básico é um fator de impacto para a proliferação do vetor da doença, pois regiões com acúmulo de água parada em sem infraestrutura adequada são essenciais para a reprodução do mosquito Aedes aegypti (Ministério da Saúde, 2024).

O modelo indicou que há 63% de variabilidade no risco de contaminação por dengue a partir dos indicadores socioeconômicos dos municípios. O Pseudo R^2, de valor acima de 0,6, é considerado satisfatório em estudos aplicados de epidemiologia social, indicando que a variação da probabilidade de contaminação é coerente (Menard, 2002).

Resultados da Regressão Logística GLM Binomial para os municípios de Goiás	
Métrica	**Valor**
Variável Dependente	CONTAMINACAO_INDICE
Modelo	GLM, Binomial, função Logit
Número de Observações	246 municípios
Graus de liberdade dos resíduos	243
Número de parâmetros estimados	2
Log-Likelihood	-22,018
Deviance	44,036
Pearson qui-quadrado	66,2
Pseudo R^2	0,6321

Tabela 2: Tabela de resultados da regressão logística binomial aplicada para os municípios do estado de Goiás.

A interação DENGUE:TOTAL, embora tenha um coeficiente negativo muito pequeno é estatisticamente significativo. Para populações maiores o impacto da quantidade de casos de contaminação de dengue na probabilidade de contaminação é reduzido, ou seja, a influência dos casos diminui à medida que o município tem sua população aumentada.

Termo	**Coeficiente**	**Erro Padrão**	**Estatística Z**	**P-valor**	**IC 95% inferior**	**IC 95% superior**
Intercepto	-5,4657	0,870	-6,281	<0,001	-7,171	-3,760
DENGUE	0,0819	0,014	5,698	<0,001	0,054	0,110
DENGUE: TOTAL	-6,196e-08	1,1e-08	-5,565	<0,001	-8,38e-08	-4,01e-08

Tabela 3: Tabela de coeficientes da regressão logística binomial aplicada para os municípios do estado de Goiás.

A matriz de correlação foi elaborada por meio de índices atribuídos a cada variável. Todas as variáveis estão disponíveis no apêndice. O índice F1 representa a população total de cada município, enquanto os índices de F2 a F34 representam os indicadores socioeconômicos. O índice F33 representa a variável quantitativa dos casos de dengue por município, e o índice F34 representa a variável dicotômica CONTAMINACAO_INDICE.

Par de Variáveis	**Coeficiente de correlação**	**Intensidade da correlação**	**Direção da correlação**	**Interpretação**
F1 e F2	0,18	Fraca	Positiva	Tendência pequena de que quando F1 aumenta, F2 aumenta levemente.
F3 e F4	0,27	Fraca a Moderada	Positiva	Tendência levemente forte de que quando F3 aumenta, F4 também aumenta.
F5 e F6	0,11	Muito Fraca	Positiva	Correlação quase desprezível.
F7 e F8	0,35	Moderada	Positiva	F7 e F8 tendem a sofrer alterações juntas.
F9 e F10	-0,04	Muito Fraca	Negativa	Correlação quase desprezível.
F11 e F12	0,01	Praticamente nula	Positiva	Correlação praticamente nula.
F13 e F14	0,09	Muito Fraca	Positiva	Correlação muito fraca.
F15 e F16	0,05	Muito Fraca	Positiva	Correlação muito fraca.
F17 e F18	-0,13	Fraca	Negativa	Pequena tendência de que quando F17 aumenta, F18 diminui.

F19 e F20	0,02	Praticamente nula	Positiva	Correlação praticamente nula.
F21 e F22	-0,13	Fraca	Negativa	Pequena tendência de que quando F21 aumenta, F22 diminui.
F23 e F24	-0,05	Muito Fraca	Negativa	Correlação muito fraca.
F25 e F26	-0,16	Fraca	Negativa	Pequena tendência de que quando F25 aumenta, F26 diminui.
F27 e F28	-0,03	Muito Fraca	Negativa	Correlação quase desprezível.
F29 e F30	0,04	Muito Fraca	Positiva	Correlação muito fraca.
F31 e F32	0,09	Muito Fraca	Positiva	Correlação muito fraca.
F33 e F34	0,26	Fraca a Moderada	Positiva	Tendência levemente forte de que quando F33 aumenta, F34 também aumenta.

Tabela 4: Tabela descritiva das variáveis correlacionadas, obtida pela matriz de correlação.

Para os municípios de todas as unidades federativas, incluindo o estado de Goiás e excluindo o distrito federal, os resultados não indicaram uma relação tão aderente ao proposto.

A ausência de abastecimento de água tratada, coleta de resíduos e drenagem de águas pluviais propicia ambientes ideais para a reprodução e proliferação do Aedes aegypti (Ministério da Saúde, 2022). Ainda, municípios com infraestrutura sanitária precária possuem incidência de contaminação de dengue até três vezes maior do que em regiões dotadas de serviços de saneamento consolidados (Teixeira et al., 2009).

A renda média mensal de uma população se apresenta como um forte indicador de risco de contaminação, estando associados à maior vulnerabilidade social quando as condições de moradia são inadequadas e ao acesso reduzido a serviços de saúde e

educação para prevenção de doenças (Donalísio & Freitas, 2015). Diante deste contexto, é importante atribuir pesos adicionais às variáveis relacionadas à saneamento e renda média ao determinar a variável dependente do modelo. O acréscimo entre 10% e 20% aos pesos tem como objetivo refletir a influência estrutural desproporcional que essas variáveis exercem sobre a dinâmica de transmissão da dengue, pois a consideração uniforme entre todas as variáveis independentes não representa de forma adequada o risco real.

Os seguintes pesos foram atribuídos às variáveis relacionadas aos domicílios ligados à rede sanitária geral:

Variável	**Peso**
Proporção_de_crianças_de_0_a_5_anos_residentes_em_domicilios _particulares_permanentes_com_saneamento_inadequado	+10%
Proporção_de_crianças_de_0_a_5_anos_residentes_em_ domicilios_particulares_permanentes_com_responsável_ ou_cônjuge_analfabeto_e_saneamento_inadequado	+10%
Proporção de domicilios particulares permanentes por tipo de saneamento Adequado	+10%
Proporção de domicilios particulares permanentes por tipo de saneamento Semi-Adequado	+15%
Proporção de domicilios particulares permanentes por tipo de saneamento Inadequado	+20%
População residente em domicilios particulares permanentes com saneamento inadequado e rendimento mensal total domiciliar per capita nominal Total	+10%
População residente em domicilios particulares permanentes com saneamento inadequado e rendimento mensal total domiciliar per capita nominal de Até R$70	+15%
População residente em domicilios particulares permanentes com saneamento inadequado e rendimento mensal total domiciliar per capita nominal de Até 1/4 SM (=R$128)	+10%
População residente em domicilios particulares permanentes com saneamento inadequado e rendimento mensal total domiciliar per capita nominal de \nAté 1/2 SM (=R$255)	+10%
População residente em domicilios particulares permanentes com saneamento inadequado e rendimento mensal total domiciliar per capita nominal de Até 60% da mediana (=R$225)	+10%

Tabela 5: Tabela de Variáveis cujo peso foi atribuído para gerar o índice CONTAMINACAO_INDICE pela média ponderada.

No modelo GLM (Generalized Linear Model), considerando a variável independente TOTAL – população total de cada município do Brasil – os resultados foram mais distantes dos resultados obtidos com o modelo que abrange somente o estado de Goiás.

Com 5564 observações, conforme indicado na figura 7 no Apêndice, o coeficiente da variável TOTAL indica que no aumento de 1 indivíduo contaminado por município aumentaria os log-odds em $2.77e^{-05}$ unidades. O p-valor da variável DENGUE é menor que 0.001, muito abaixo de 0.05, indicando que estatisticamente, o modelo evidencia associação com a variável dependente. O Pseudo-R² no valor aproximado de 0.11 indica que o modelo apresenta um ajuste bom aos dados. O Log-Likelihood no valor aproximado de -3524,4 não é capaz de indicar individualmente se o modelo apresenta uma probabilidade de acerto relativamente grande, por se aproximar de zero.

A representação do modelo logístico binário, utilizando os índices CONTAMINACAO_INDICE e TOTAL, adicionando os valores previstos de probabilidade na variável phat, em uma predição para cada 100000 (cem mil) habitantes em um município resultou na curva sigmóide representando a relação entre a probabilidade predita phat e a variável dependente CONTAMINACAO_INDICE para os municípios do país todo, conforme apresentado na figura 8 no Apêndice. Os pontos laranjas representam as observações reais e a curva roxa representa as estimativas do modelo. No caso de maiores concentrações de probabilidades baixas para casos negativos e altas para positivos, o modelo apresenta bom poder discriminativo mesmo existindo sobreposição entre grupos.

A matriz de confusão, presente na figura 9 no Apêndice, demonstra que para um cutoff de 0.4, 1739 casos de contaminação foram verdadeiros positivos, 1172 falsos positivos, 852 falsos negativos e 1801 verdadeiros negativos, indicando que o modelo é razoável em termos de identificar corretamente os casos de contaminação.

O gráfico que apresenta a variação da especificidade e da sensitividade em função do cutoff, onde a sensitividade representa a capacidade do modelo em identificar corretamente os verdadeiros positivos e a especificidade representa a capacidade do modelo de identificar corretamente os verdadeiros negativos, indica que para um cutoff de 0.4 o modelo apresenta um bom equilíbrio, conforme apresentado na figura 10 no Apêndice.

Curva ROC	**AUC**	**Descrição**

Variáveis independentes do modelo de regressão contemplando somente os municípios do estado de Goiás	0,92	O modelo teve um desempenho excelente para distinguir entre contaminado e não contaminado
Sensitividade e especificidade para os municípios de todas as unidades federativas, exceto o distrito federal	0,706	O modelo possui um bom desempenho, com 70,6% de chance de classificar um par de amostras corretamente.
Previsão de contaminação para uma população de 1000000 habitantes no modelo stepwise	0,71	O modelo possui um desempenho razoável para distinguir casos de contaminação e não contaminação.
Hiperparâmetros ajustados (min_samples_leaf=2, min_samples_split=5, n_estimators=200) utilizando Random Forest na base de treino.	0,805	O modelo se mostrou capaz de discriminar entre municípios de alto e baixo risco de contaminação.

Tabela 6: Tabela de desempenhos das curvas ROC em diferentes cenários, datasets e previsões.

O modelo logístico binário pela função sm.Logit.from_formula, do tipo MLE - Maximum Likelihood Estimation - apresentou valores idênticos ao modelo anterior - GLM - que possui o Pseudo R² ligeiramente maior, indicando um bom ajuste, conforme apresentado na figura 12 no Apêndice. O teste de razão de verossimilhança indica que o modelo é significativo.

$$LLR = 2\left(LogLik_{modelo} - LogLik_{nulo}\right) = 2(-3524,4 + 3843,5) = 638,2$$

O método de estimativa Maximum Likelihood Estimation, com Pseudo R² de 0,083 e com LogLikelihood de -3524,4 e um LL-Null de -3843,5, indicou que há um efeito leve porém significativo da variável TOTAL em função das demais variáveis independentes sobre o impacto na contaminação de dengue por município, o que indica que quanto maior a população total do município, maior a probabilidade de um índice de contaminação mais elevado.

Após o ajuste dos hiperparâmetros e a utilização do Random Forest, o modelo foi treinado e avaliado em um conjunto de validação. A acurácia resultou em 72% representando a proporção de predições corretas em relação ao total de predições realizadas. A área sob a curva ROC (AUC) resultou em 0,805, o que reflete a capacidade do

modelo de discriminar entre municípios de alto e baixo risco de contaminação, sendo considerado um valor bom para tarefas de classificação, por estar acima de 0,8.

Métrica	**Não contaminação**	**Contaminação**	**Média Ponderada**
Precisão	0,71	0,74	0,72
Recall	0,81	0,62	0,72
F1-Score	0,76	0,67	0,72
Acurácia			0,72

Tabela 7: Tabela de desempenho do modelo Random Forest.

A importância das variáveis indicou que fatores relacionados a condições de vida e infraestrutura urbana exerceram maior influência nas predições do modelo. A tabela 8 apresenta as variáveis de maior importância que se destacaram.

Variáveis do Modelo	**Grau de Importância**
população total	0,088
População residente em domicílios particulares permanentes	0,083
Percentual da população com acesso à rede de esgoto adequado	0,053
Percentual da população com acesso à rede de esgoto semi-adequado	0,049
Percentual da população com acesso à rede de esgoto inadequado	0,035
Rendimento médio domiciliar per capita de menos de um salário mínimo	0,036
Domicílios particulares com responsável ou cônjuge analfabeto	0,051

Tabela 8: Tabela de Variáveis de importância no modelo.

Esses resultados corroboram com a hipótese de que a vulnerabilidade à contaminação de dengue não é apenas uma questão climática, e que pode estar fortemente associada a fatores socioeconômicos estruturais. Tais resultados verificam a viabilidade do modelo proposto de predição do risco de contaminação por dengue com base em variáveis socioeconômicas e demográficas.

No entanto, apesar dos resultados satisfatórios, o modelo possui algumas limitações como a ausência de variáveis ambientais como precipitação e temperatura média, que pode limitar a capacidade do modelo em capturar fatores sazonais de proliferação do Aedes Aegypti, vetor da arbovirose, e o desequilíbrio entre municípios altamente contaminados e municípios com baixa contaminação pode influenciar o treinamento, mesmo que a validação cruzada minimize o efeito.

Conclusão

De maneira geral, as variáveis socioeconômicas utilizadas no modelo como população total, proporção de indivíduos residentes em moradia com saneamento inadequado, semi-adequado, adequado e demais indicadores de analfabetismo e de vulnerabilidade social, têm um impacto relativo sobre a probabilidade de contaminação de um indivíduo com dengue. Com base nas análises realizadas dos gráficos e resultados, há uma relação entre a situação de vulnerabilidade social do indivíduo e a probabilidade de contrair dengue e a presença de condições vulneráveis de saneamento e renda per capita mais baixa que indica uma maior chance de contaminação da doença.

Foi constatado que os indicadores socioeconômicos de um município possuem influência significativa sobre a probabilidade de indivíduos de sua população se contaminarem com dengue. O desempenho obtido no conjunto de teste, avaliado pela métrica de ROC AUC, demonstrou que as variáveis como renda média, acesso à infraestrutura de saneamento básica, índice de escolaridade e condições habitacionais estão associadas ao risco de contaminação por dengue. Municípios com piores indicadores socioeconômicos apresentaram maior probabilidade de incidência da doença, o que sugere que a vulnerabilidade social é um fator crítico no contexto epidemiológico da dengue. Esses resultados corroboram a literatura existente, apontando a precariedade social como um vetor de intensificação de doenças transmitidas por arbovírus, como a dengue, e reforçam a necessidade de políticas públicas que integrem ações de combate à desigualdade e de promoção da saúde coletiva.

Referências

1. Análise espacial da dengue e o contexto socioeconômico no município do Rio de Janeiro, RJ, Almeida S., Medronho A. e Valencia O. 2009. https://www.scielo.br/j/rsp/a/d7KJxrZX4H7x597ZDBsdKrJ/abstract/?lang=en#ModalTutors

2. Brasil é país com mais casos de dengue no mundo https://agenciabrasil.ebc.com.br/saude/noticia/2023-12/brasil-e-pais-com-mais-casos-de-dengue-no-mundo-mostra-dados-da-oms

3. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. 2013. Applied Logistic Regression. 3rd Edition. John Wiley & Sons. Hoboken, New Jersey, United States of America. Disponível em: https://dl.icdst.org/pdfs/files4/7751d268eb7358d3ca5bd88968d9227a.pdf. Acesso em: 06 de novembro de 2024.

4. IBGE https://www.ibge.gov.br/estatisticas/sociais/populacao/9221-sintese-de-indicadores-sociais.html?=&t=resultados https://www.ibge.gov.br/estatisticas/sociais/populacao/2098-np-censo-demografico/22827-censo-demografico-2022.html?edicao=35938&t=downloads

5. SES-GO DENGUE https://dados.saude.go.gov.br/dataset/dengue https://indicadores.saude.go.gov.br/public/dengue.html

6. https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON518

7. https://www.scielo.br/j/csp/a/YCXNkzcfS9LNK7JWdFvfFtB/?lang=en

8. TabNet DataSus http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/denguebac.def

9. https://censo2010.ibge.gov.br/sinopse/index.php?uf=52&dados=1

10. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

11. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

12. Fawcett, T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861–874.

13. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

14. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

15. Fawcett, T. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861–874.

16. Menard, S. (2002). *Applied Logistic Regression Analysis*. Sage.

17.Ministério da Saúde, gov.br https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/dengue

18. IBGE, 2022
https://agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/41901-censo-2022-87-da-populacao-brasileira-vive-em-areas-urbanas
19. Boletim Epidemiológico da Dengue - Rio de Janeiro - https://sistemas.saude.rj.gov.br/tabnetbd/dash/dash_dengue.htm

Apêndice

Índice da matriz de correlação	Variável
F1	POPULACAO_TOTAL
F2	PROPORCAO_DE_CRIANCAS_DE_0_A_5_ANOS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RESPONSAVEL_OU_CONJUGE_ANALFABETO
F3	PROPORCAO_DE_CRIANCAS_DE_0_A_5_ANOS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_SANEAMENTO_INADEQUADO
F4	PROPORCAO_DE_CRIANCAS_DE_0_A_5_ANOS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RESPONSAVEL_OU_CONJUGE_ANALFABETO_E_SANEAMENTO_INADEQUADO
F5	PROPORCAO_DE_DOMICILIOS_PARTICULARES_PERMANENTES_POR_TIPO_DE_SANEAMENTO_ADEQUADO
F6	PROPORCAO_DE_DOMICILIOS_PARTICULARES_PERMANENTES_POR_TIPO_DE_SANEAMENTO_SEMI-ADEQUADO
F7	PROPORCAO_DE_DOMICILIOS_PARTICULARES_PERMANENTES_POR_TIPO_DE_SANEAMENTO_INADEQUADO
F8	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_SANEAMENTO_INADEQUADO_E_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_TOTAL
F9	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES_C

	OM_SANEAMENTO_INADEQUADO_E_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_DE_ATE_R$70
F10	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_SANEAMENTO_INADEQUADO_E_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_DE_ATE_1/4_SM_(=R$128)
F11	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_SANEAMENTO_INADEQUADO_E_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_DE__ATE_1/2_SM_(=R$255)
F12	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_SANEAMENTO_INADEQUADO_E_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_DE_ATE_60%_DA_MEDIANA_(=R$225)
F13	POPULACAO_RESIDENTE_EM_DOMICILIOS_PARTICULARES_PERMANENTES
F14	PROPORCAO_DE_PESSOAS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_DE_ATE_70,00_R$
F15	PROPORCAO_DE_PESSOAS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_DE_ATE_1/4_SALARIO_MINIMO_(=_127,50_R$)
F16	PROPORCAO_DE_PESSOAS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_DE_ATE_1/2_SALARIO_MINIMO_(=_255,00_R$)
F17	PROPORCAO_DE_PESSOAS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_COM_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_DE_ATE_60%_DA_MEDIANA_-_BRASIL_

	TOTAL_(=_225,00_R$)
F18	RAZAO_ENTRE_MEDIAS_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_PESSOAS_BRANCAS/PRETAS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(A/B)
F19	RAZAO_ENTRE_MEDIAS_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_PESSOAS_BRANCAS/PARDAS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(A/C)
F20	RAZAO_ENTRE_MEDIAS_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_PESSOAS_BRANCAS/AMARELAS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(A/D)
F21	RAZAO_ENTRE_MEDIAS_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_PESSOAS_BRANCAS/INDIGENAS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(A/E)
F22	RAZAO_ENTRE_MEDIAS_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_PESSOAS_PRETAS/PARDAS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(B/C)
F23	VALOR_MEDIO_DO_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL_(R$)
F24	1_QUARTIL_DO_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL
F25	2_QUARTIL_(MEDIANA)_DO_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL
F26	3_QUARTIL_DO_RENDIMENTO_MENSAL_TOTAL_DOMICILIAR_PER_CAPITA_NOMINAL
F27	RENDIMENTO_MENSAL_TOTAL_NOMINA

	L_DE_HOMENS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(A)_VALOR_MEDIO
F28	RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_MULHERES_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(B)_VALOR_MEDIO
F29	RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_HOMENS_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(C)_MEDIANO
F30	RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_MULHERES_COM_10_ANOS_OU_MAIS_RESIDENTES_EM_DOMICILIOS_PARTICULARES_PERMANENTES_(D)_MEDIANO
F31	RAZAO_ENTRE_VALOR_MEDIO_E_MEDIANO_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_HOMENS_E_MULHERES_MEDIO_(A/B)
F32	RAZAO_ENTRE_VALOR_MEDIO_E_MEDIANO_DO_RENDIMENTO_MENSAL_TOTAL_NOMINAL_DE_HOMENS_E_MULHERES_MEDIANO_(C/D)
F33	DENGUE
F34	CONTAMINACAO_INDICE

Tabela 9: Tabela de Variáveis Independentes e Dependentes do modelo, exceto MUNICIPIO e UF.

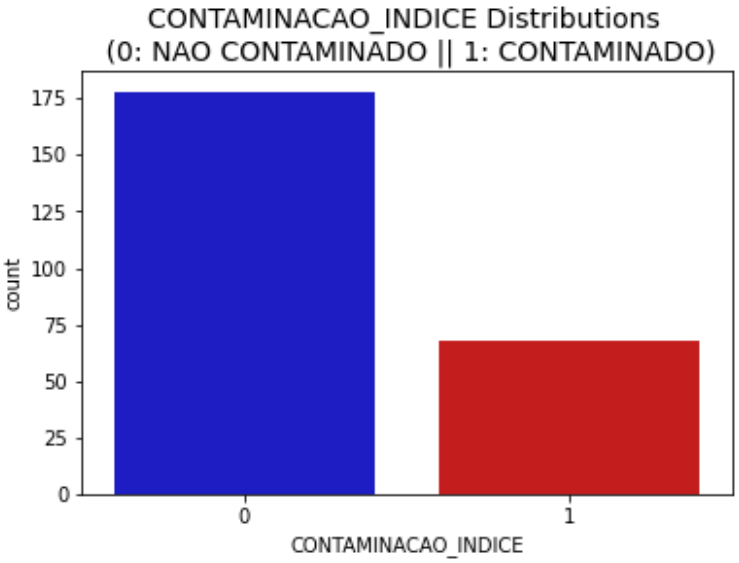

Figura 1: Distribuição de observações contendo indivíduos contaminados e não contaminados em Goiás.

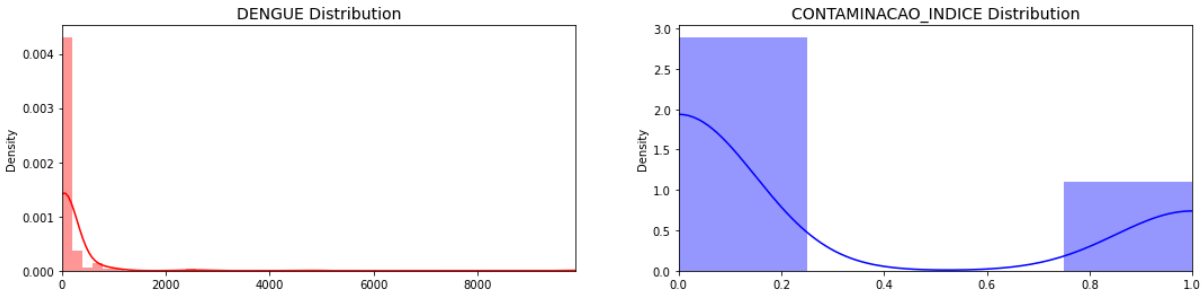

Figura 2: Densidade de distribuição de observações com indivíduos contaminados e não contaminados em Goiás.

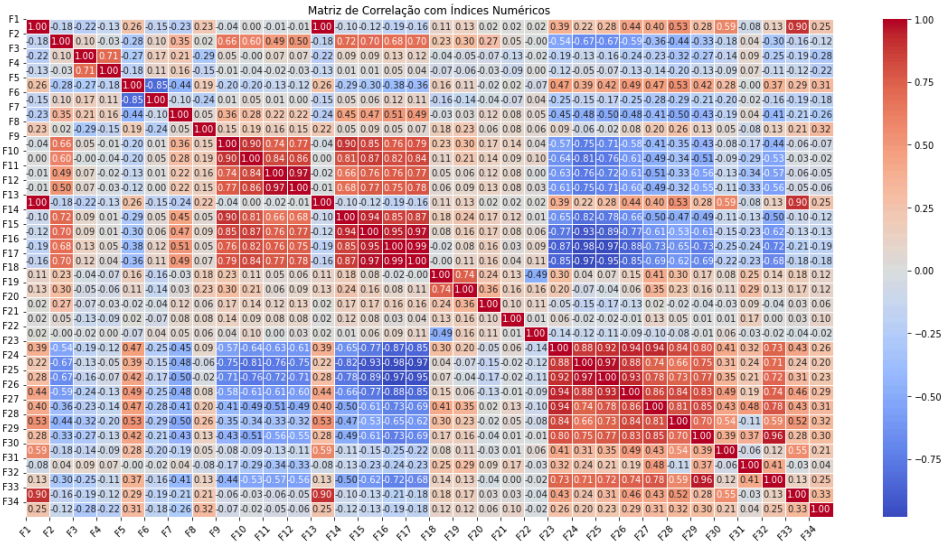

Figura 3: Matriz de correlação utilizando índices numéricos para descrever cada variável independente correspondente ao estado de Goiás.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:     CONTAMINACAO_INDICE   No. Observations:                  246
Model:                             GLM   Df Residuals:                      243
Model Family:                Binomial   Df Model:                            2
Link Function:                   Logit   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -22.018
Date:                 Mon, 03 Feb 2025   Deviance:                       44.036
Time:                         14:23:54   Pearson chi2:                     66.2
No. Iterations:                      10   Pseudo R-squ. (CS):             0.6321
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      -5.4657      0.870     -6.281      0.000      -7.171      -3.760
DENGUE          0.0819      0.014      5.698      0.000       0.054       0.110
DENGUE:TOTAL -6.196e-08   1.11e-08     -5.565      0.000   -8.38e-08   -4.01e-08
==============================================================================
"""
```

Figura 4: Resultados obtidos do modelo de Regressão Logística com a variável CONTAMINACAO_INDICE para os municípios do estado de Goiás.

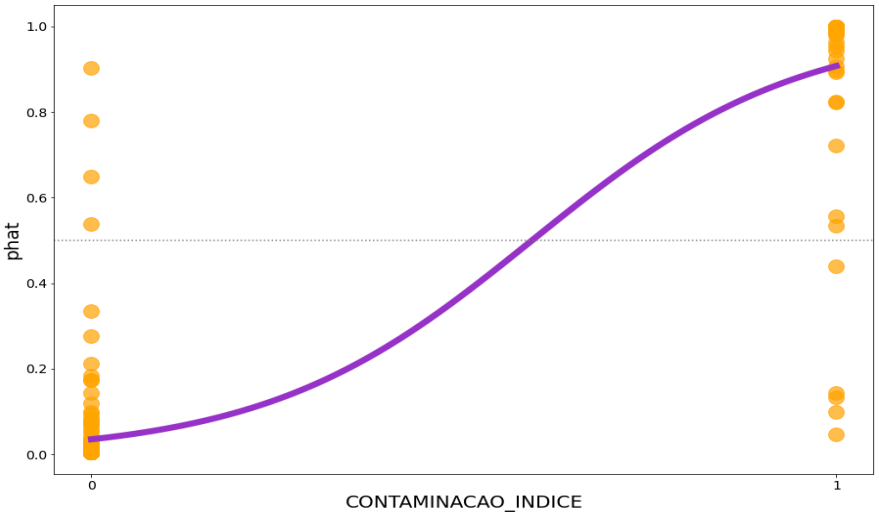

Figura 5: Curva sigmoide representando a variável que indica a probabilidade predita phat e a variável dependente CONTAMINACAO_INDICE para os municípios do estado de Goiás.

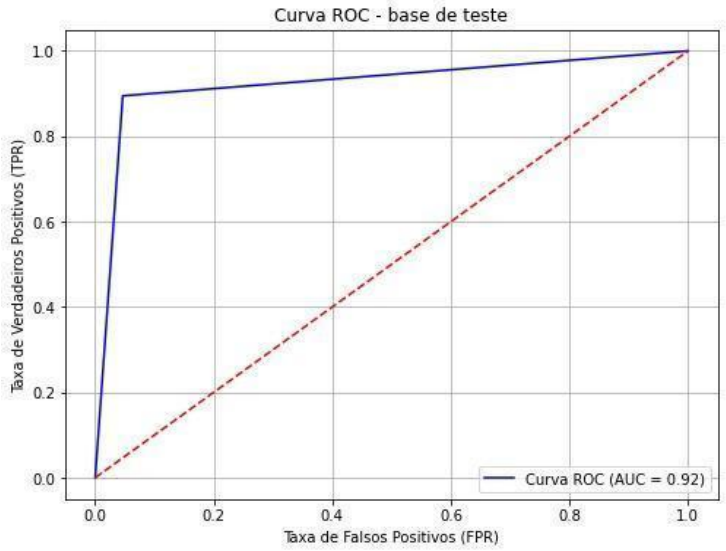

Figura 6: Curva ROC representada apenas pelas variáveis independentes.

Generalized Linear Model Regression Results

Dep. Variable:	CONTAMINACAO_INDICE	No. Observations:	5564
Model:	GLM	Df Residuals:	5562
Model Family:	Binomial	Df Model:	1
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-3524.4
Date:	Sun, 13 Apr 2025	Deviance:	7048.8
Time:	21:33:00	Pearson chi2:	8.59e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.1084
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6947	0.039	-18.037	0.000	-0.770	-0.619
TOTAL	2.77e-05	1.6e-06	17.287	0.000	2.46e-05	3.08e-05

Figura 7: Resultados obtidos do modelo Binomial GLM de Regressão Logística com a variável CONTAMINACAO_INDICE para os municípios de todas as unidades federativas, exceto o distrito federal.

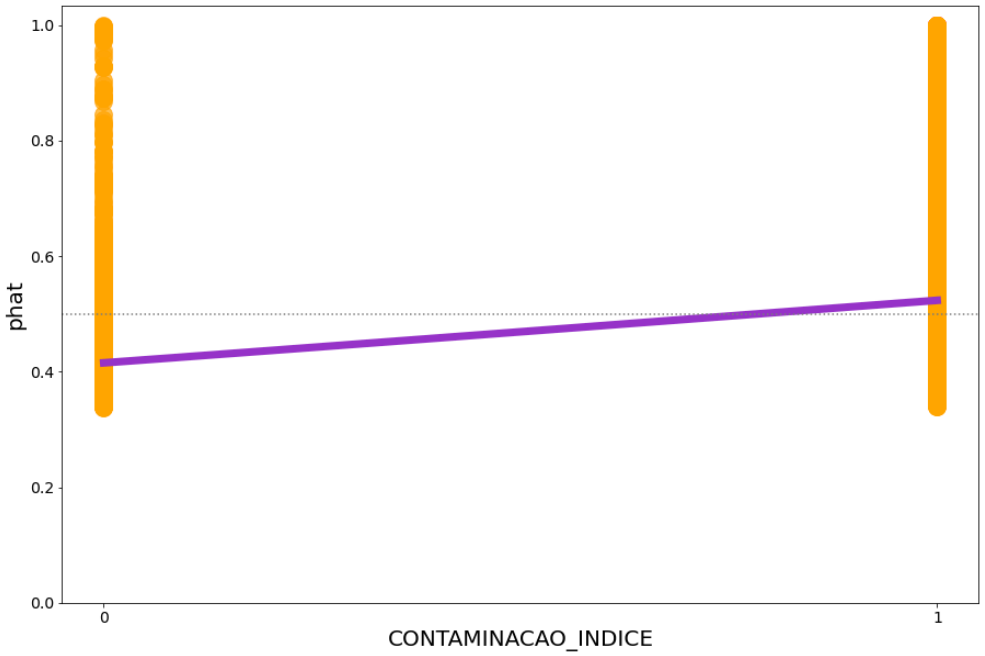

Figura 8: Curva sigmoide representando a variável que indica a probabilidade predita phat e a variável dependente CONTAMINACAO_INDICE para os municípios de todas as unidades federativas, exceto o distrito federal.

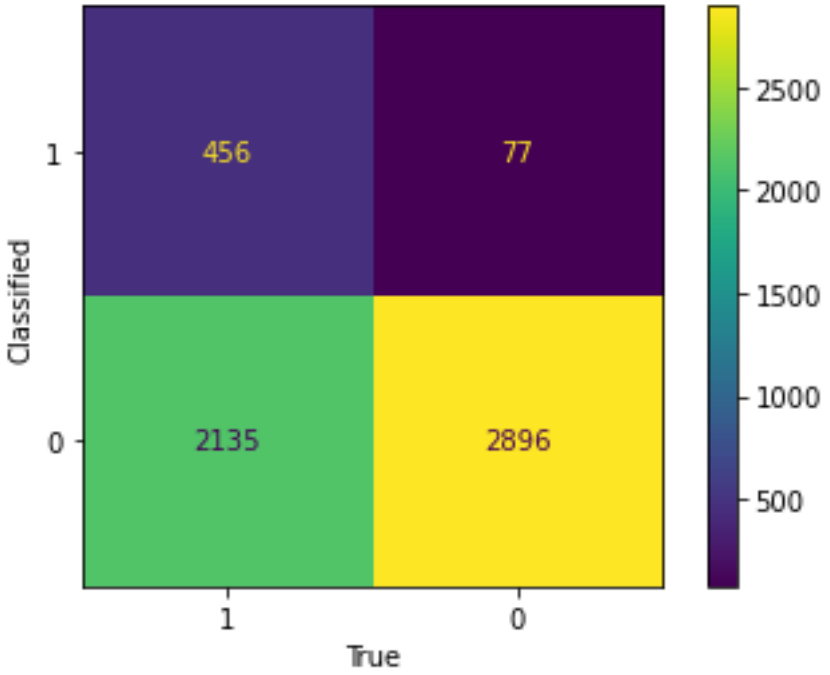

Figura 9: Matriz de confusão representando a variável que indica a probabilidade predita phat e a variável dependente CONTAMINACAO_INDICE para os municípios de todas as unidades federativas, exceto o distrito federal.

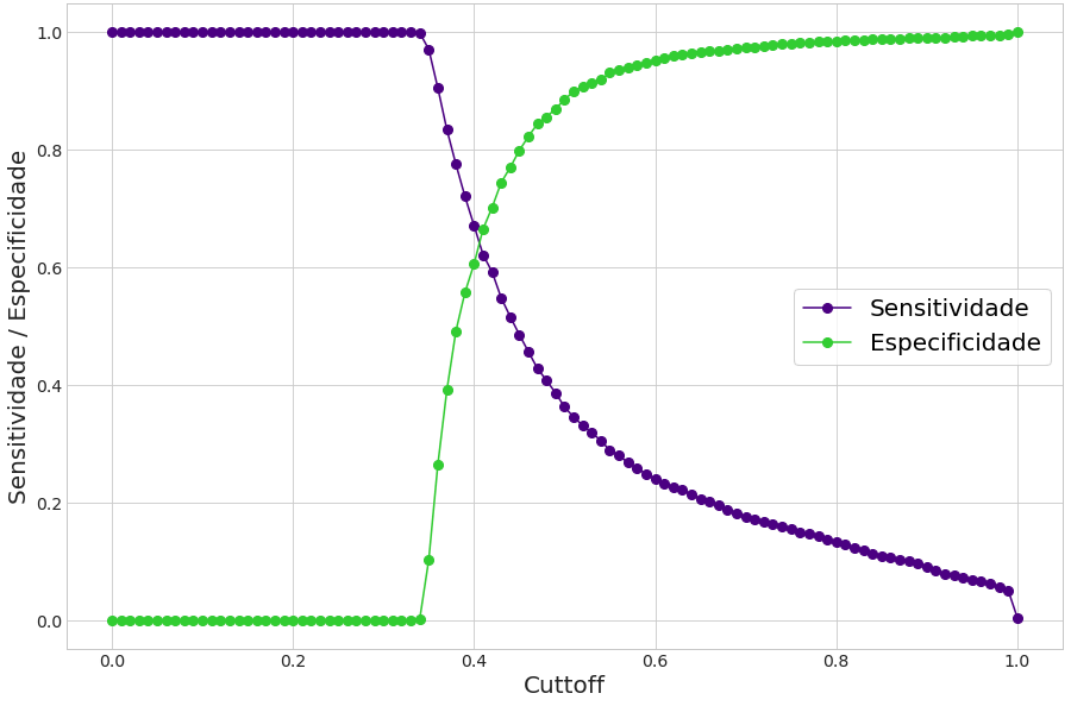

Figura 10: Gráfico da variação da especificidade e da sensitividade em função do cutoff representando a variável que indica a probabilidade predita phat e a variável dependente CONTAMINACAO_INDICE para os municípios de todas as unidades federativas, exceto o distrito federal.

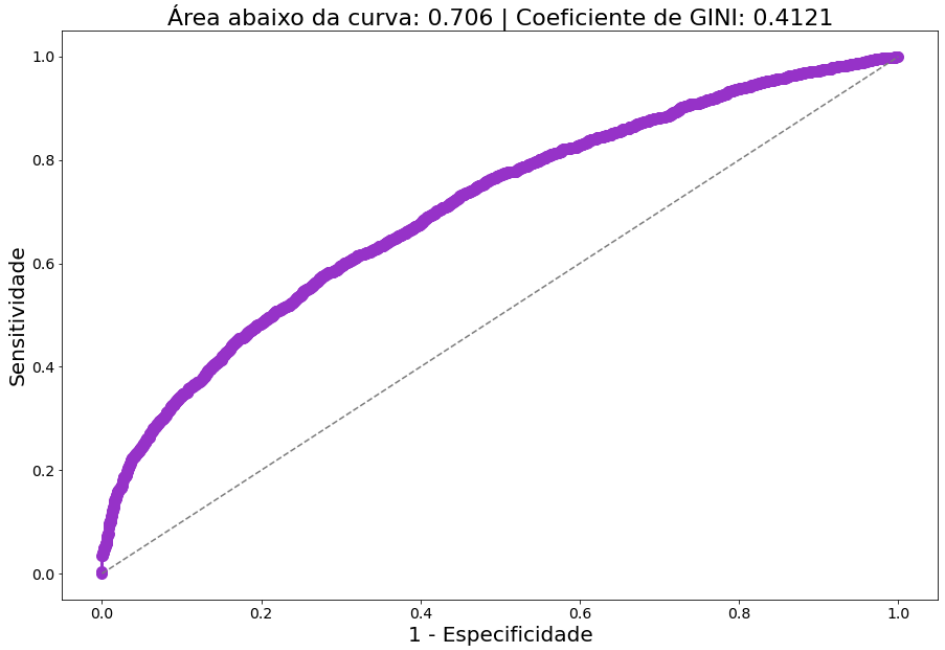

Figura 11: Curva ROC do pacote metrics representando a sensitividade e especificidade para os municípios de todas as unidades federativas, exceto o distrito federal.

Logit Regression Results

Dep. Variable:	CONTAMINACAO_INDICE	No. Observations:	5564
Model:	Logit	Df Residuals:	5562
Method:	MLE	Df Model:	1
Date:	Mon, 14 Apr 2025	Pseudo R-squ.:	0.08304
Time:	20:38:25	Log-Likelihood:	-3524.4
converged:	True	LL-Null:	-3843.5
Covariance Type:	nonrobust	LLR p-value:	7.732e-141

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-0.6947	0.039	-18.037	0.000	-0.770	-0.619
TOTAL	2.77e-05	1.6e-06	17.287	0.000	2.46e-05	3.08e-05

Figura 12: Resultados obtidos do modelo Binomial GLM de Regressão Logística com a variável CONTAMINACAO_INDICE para os municípios de todas as unidades federativas, exceto o distrito federal.

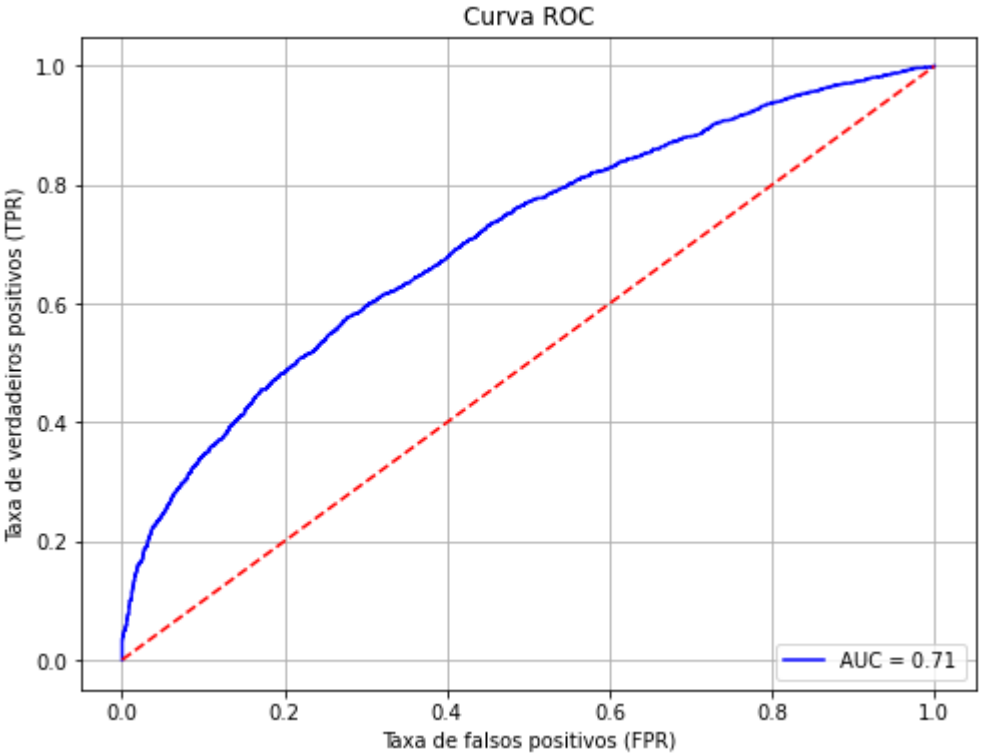

Figura 13: Curva ROC representada pela variavel phat do modelo e a variável dependente CONTAMINACAO_INDICE.

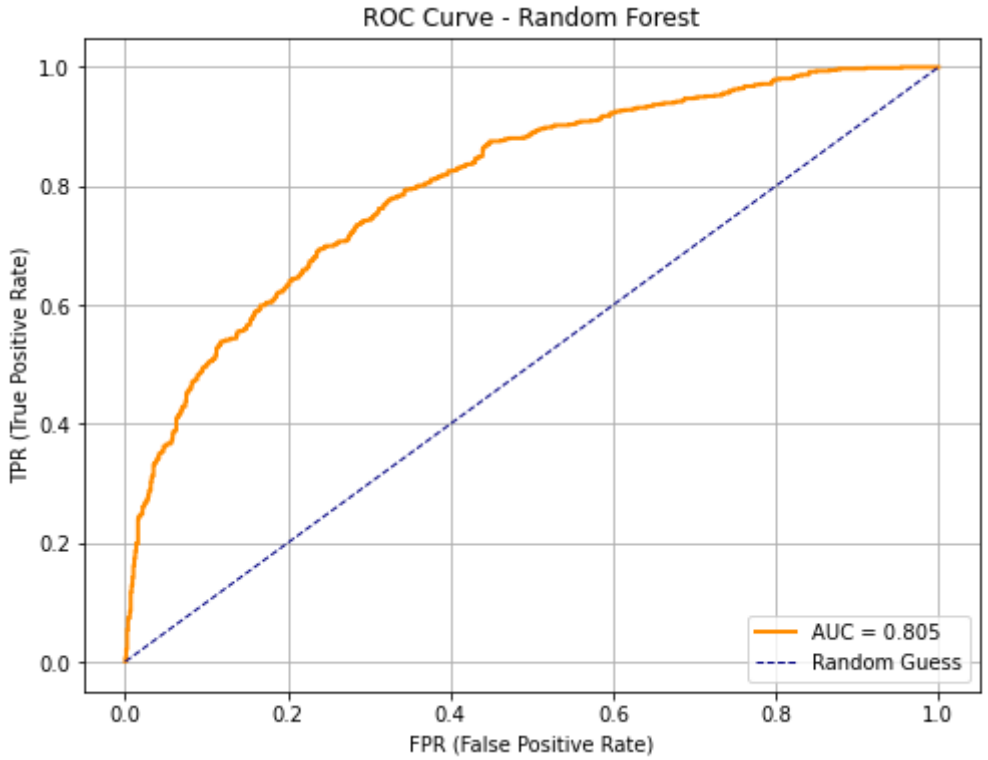

Figura 14: Curva ROC resultante do modelo treinado por Random Forest.