Projeto de Graduação

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Dezembro 2023

Predição semanal de casos de dengue no estado do Rio de Janeiro.

Guilherme Faveret Garcia de Souza

PUC RIO

www.ele.puc-rio.br

Projeto de Graduação

DEPARTAMENTO
DE ENGENHARIA
ELÉTRICA

Predição semanal de casos de dengue no estado do Rio de Janeiro

Aluno: Guilherme Faveret Garcia de Souza

Orientador: Hélio Cortês Vieira Lopes

Trabalho apresentado com requisito parcial à conclusão do curso de Engenharia Elétrica na Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil.

Agradecimentos

Agradeço ao meus pais por terem me incentivado a seguir a trilha dos estudos e a todos meus professores, que foram imprescindíveis nesse caminho. Sobretudo, gostaria de agradecer ao meu orientador pelo apoio neste projeto e ao professor Silvio Hamacher, que me introduziu aos estudos de dados.

Não poderia deixar de dar os devidos méritos a todos os meus amigos que fizeram essa trajetória até a formatura bem mais aprazível, seja dentro ou fora da universidade.

DEPARTAMENTO
DE ENGENHARIA
ELÉTRICA

Resumo

Com o aumento do número de casos de dengue no estado do Rio de Janeiro em 2023, faz-senecessário elevar a vigilância nesses casos e direcionar recursos corretamente. Nesse sentido, o trabalho se propõe a construir um modelo preditivo de casos de dengue para os municípios do Rio de Janeiro, utilizando variáveis climáticas e socio-demográficas. Para isso, são utilizadas técnica de ciência de dados, estatística e computação para analisar uma base 5 anos de casos de dengue por semana epidemiológica em cada municípios do RJ, enriquecendo-a com novas variáveis explicativas e construindo um modelo preditivo através de algoritmos de aprendizado de máquina.

Palavras-chave: Dengue; Estatística; Aprendizado de máquina; Predição

DEE DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Weekly Prediction of Dengue Cases in the State of RJ (Rio de Janeiro)

Abstract

With the increase in the number of dengue cases in the state of Rio de Janeiro in 2023, it becomes necessary to enhance surveillance of these cases and properly allocate resources. In this context, this study aims to build a predictive model for dengue cases in the municipalities of Rio de Janeiro, using climatic and socio-demographic variables. To this end, data science techniques, statistics, and computing are used to analyze a database of dengue cases over five years by epidemiological week in each municipality of RJ, enriching it with new explanatory variables and constructing a predictive model using machine learning algorithms.

Keywords: Dengue; Statistics; Machine Learning; Prediction

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Dengue é uma enfermidade viral, propagada pelo mosquito Aedes aegypti, e constitui um desafio significativo para a saúde pública no Brasil e em outras regiões tropicais. Reintroduzida no Brasil em 1981, a doença tem se tornado progressivamente mais comum, desencadeando várias epidemias. Durante esses surtos, observou-se alterações no perfil epidemiológico, reemergência de distintos sorotipos e um aumento nas notificações de casos severos e fatais.

O primeiro surto registrado de dengue no Brasil aconteceu em 1981, desencadeado pelos sorotipos DENV-1 e DENV-2, no estado de Roraima. Em 1986, o Rio de Janeiro foi atingido por uma epidemia que posteriormente se espalhou para outros estados do Sudeste. Quase 15 anos depois, em 2000, um novo sorotipo, DENV-3, foi identificado (NUNES et al., 2019).

Na década seguinte, novos surtos ocorreram com a reemergência dos sorotipos DENV-1 e DENV-2. Em 2010, foi registrado o maior número de óbitos relacionados à dengue até aquela data (NUNES et al., 2018).

Recentemente, em 2023, houve um alarmante aumento de casos de dengue no continente americano, o mais elevado já registrado. Até agosto de 2023, mais de 3 milhões de casos foram reportados, incluindo 28 mil casos graves e 1.800 mortes, com o Brasil sendo o país mais afetado. As regiões Sul, Sudeste e Centro-Oeste foram as mais impactadas, com 1.530.940 casos confirmados laboratorialmente (e 946 mortes) até o início de setembro, superando os 1.313.805 casos do período correspondente no ano anterior. Considerando as notificações não confirmadas laboratorialmente, o número total de casos em 2023 poderia ultrapassar 2 milhões.

Em 2023, os sorotipos DENV-1 e DENV-2 foram os mais prevalentes, mas também foram reportados casos do sorotipo DENV-3 no Norte do país e em outras nações da América Latina, sugerindo a possibilidade de uma futura epidemia desse sorotipo no Brasil (PASQUINI, 2023).

Nesse sentido, vale ressaltar como funciona a imunização à doença. A dengue possui 4 sorotipos. Quando um indivíduo é contaminado com um deles, ele se torna imune apenas a este sorotipo, mas ainda pode ser infectado pelos demais (PASQUINI, 2023). Se a reincidência acontecer, aumenta-se a probabilidade de se tornar um caso grave.

Como não há epidemias do sorotipo 3 no Brasil há mais de 15 anos, boa parte da população não está imunizada, porém podem recentemente ter sido contaminados por outro sorotipo. Portanto, um cenário alarmante com mais casos graves não seria improvável.

Tendo isso em vista, é de extrema importância estabelecer mecanismos efetivos de vigilância e prevenção de dengue. Nesse intuito, diversas abordagens científicas têm sido feitas, como visitas de campo para identificar e eliminar focos de dengue, preparação de médicos e enfermeiros para identificação de casos graves de dengue rapidamente, além de campanhas de vacinação e conscientização contra a dengue (LESSA, 2014).

Complementarmente a isso, tem sido desenvolvidas diversas soluções que utilizam ciência de dados para previsão e monitoramento de casos.

Uma solução interessante é a ferramenta de previsão da quantidade de mosquitos desenvolvida pela OFF! e o Google Cloud (REVISTA CRESCER, 2023). Essa ferramenta, que utiliza dados climáticos e informações sobre o ciclo de vida do mosquito, pode prever a frequência de mosquitos em determinadas regiões com até sete dias de antecedência

Complementarmente, pesquisadores brasileiros desenvolveram um método baseado em inteligência artificial que utiliza dados de internações hospitalares e o número de casos confirmados de dengue, além de informações coletadas de armadilhas de ovos de mosquitos. Este método, validado cientificamente, consegue prever surtos de dengue com seis semanas de antecedência e tem uma precisão superior a 90% (SANCHEZ-GENDRIZ et al., 2022).

Pode-se citar ainda outros estudos desenvolvidos com este propósito. Uma tese da Universidade de São Paulo utilizou modelos de aprendizado de máquina para não só prever a prevalência da doença, como também para compreender os fatores propulsores, assim possibilitando intervenções eficazes. (ROSTER, 2023).

Entender tais fatores é essencial para que políticas eficientes de combate à dengue sejam implementadas e cada vez mais se sabe que não só fatores epidemiológicos contribuem para o surgimento de novos casos de dengue, como também questões socioambientais. Hoje, há o entendimento de que a dengue se relaciona fortemente à fatores como, por exemplo, o crescimento urbano desordenado, desigualdade social, moradias precárias e deficiências de fornecimento de água (BEZERRA e MATOS, 2023).

Ademais, o impacto das características climáticas no desenvolvimento da dengue é um tópico amplamente estudado. A chuva, umidade e temperatura são fatores cruciais que influenciam tanto no ciclo de vida do mosquito Aedes aegypti quanto na propagação do vírus da dengue. Estudos demonstram que mudanças climáticas, como alterações na temperatura, influenciam no desenvolvimento, reprodução e comportamento dos mosquitos, afetando consequentemente a transmissão da doença.

No contexto do Rio de Janeiro, a combinação de clima favorável e urbanização desordenada, que muitas vezes resulta em acúmulo de lixo e água parada, cria um ambiente ideal para a proliferação do mosquito vetor. Estratégias eficazes de controle da dengue devem considerar esses fatores ambientais e climáticos, além de envolver ações de prevenção, vigilância, diagnóstico precoce e tratamento adequado.

O controle efetivo da dengue requer o desenvolvimento de estratégias integradas e sustentáveis, envolvendo ações de prevenção, vigilância, diagnóstico precoce e tratamento adequado dos casos, além do engajamento da população e de diferentes setores da sociedade. Tais modelos de previsão ajudam a identificar áreas e períodos de maior risco, permitindo uma alocação mais eficaz de recursos e esforços de controle.

Tendo isso em vista, O intuito deste trabalho é apoiar na vigilância e prevenção de casos de dengue, se baseando em fatores epidemiológicos, climáticos e socioambientais. Reconhecendo as diferenças socioculturais, climáticas e geográficas de cada município, o modelo construído leva em conta as singularidades de cada local. Desta forma, espera-se que o modelo seja capaz de estimar com precisão suficiente a quantidade de casos futuros de dengue por semana epidemiológica e município.

Desta forma, espera-se que o modelo seja capaz de estimar com precisão suficiente a quantidade de casos futuros de dengue por semana epidemiológica e município.

2. Revisão Teórica

a. Estatística

Estatística se define como uma ciência que se propõe a obter, organizar, resumir, analisar e interpretar dados através de um conjunto de métodos. Sendo assim, esse ramo auxilia na extração de conhecimento a partir da interpretação de seu significado.

Dentro desta ciência, uma série de métodos são aplicados para sumarizar o comportamento do conjunto de dados sendo analisado. Esta é o que se chama de estatística descritiva e contém definições e medidas primordiais para um entendimento geral dos dados.

Medidas de tendência central objetivam sintetizar em um único número os dados observados, dando a ideia de valor médio dos dados. A mais conhecida dessas medidas certamente é a média, que é calculada pela soma dos valores das observações dividida pela quantidade de observações. Se um dado possui n observações, então a média $\bar{x}$ é obtida pela fórmula abaixo.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \quad (1)$$

Já a mediana é o valor que divide o conjunto de dados em duas partes iguais (com mesmo número de observações) após a ordenação dos dados. Assim, essa medida dá uma noção de posição central ordenada e é menos sensível à outliers do que a média.

Na mesma lógica da mediana, que divide os dados em dois grupos de igual tamanho, os quartis dividem o conjunto em quatro partes ordenadas iguais (DOS SANTOS, 2023). Ou seja, o 1º quartil (Q1 ou quartil de 25%) é o valor abaixo do qual está 25% das observações; o 2º quartil (50%) divide em partes de igual tamanho e equivale à mediana; e o 3º quartil (Q3 ou 75%) é o valor abaixo do qual está 75% das observações conforme pode ser visto na Figura 1 abaixo.

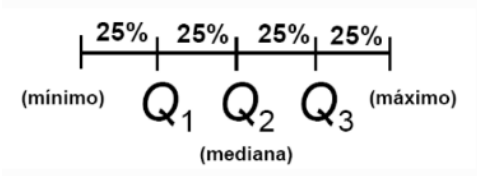

Figura 1- Representação de quartis
Fonte: CRUZ, 2022

Uma medida derivada dos quartis é o intervalo interquartil (IQR), a qual denota a diferença entre os valores de Q3 e o Q1. A partir dele, podemos calcular os limites superior (2) e inferior (3) dos dados, que são extremamente importantes para a definição de outliers, que são as observações situada fora desses limites definidos. A fórmula usada para calcular cada um desses limites pode ser vista abaixo.

$$Limite\ superior = Q3 + 1{,}5IQR \quad (2)$$

$$Limite\ inferior = Q1 - 1{,}5IQR \quad (3)$$

A partir desses conceitos, surgem gráficos como o boxplot, que representa cada um dos quartis e os limites inferior e superior dos valores das observações. Nesse gráfico, os outliers são denotados por pontos acima e abaixo dos limites superior e inferior, respectivamente. A Figura 2 ilustra alguns conceitos do Boxplot.

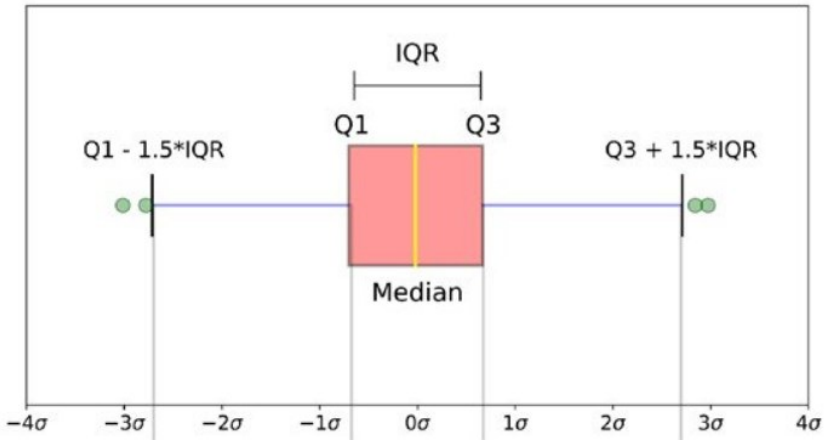

Figura 2 - Interpretação de um Boxplot
Fonte: MACIEL, 2023

Existe ainda uma terceira e menos utilizada medida de tendência central denominada moda. Essa medida representa o valor mais frequente dos dados analisados (DOS SANTOS, 2023). Vale notar que um conjunto de dados pode possuir mais de uma moda, caso exista mais de um valor igualmente frequente.

Por mais importantes que as medidas de tendência central sejam, elas não conseguem representar como um todo o comportamento dos dados. Existe ainda um outro aspecto muito importante a ser analisado que á variabilidade e um conjunto de métodos estatísticos se propõe a resumir essa informação através de outras medidas.

A primeira e principal dessas medidas de variabilidade é a variância (4). Essa medida representa a dispersão dos valores em relação à média e quanto maior seu valor, maior é a variabilidade dos dados (DOS SANTOS, 2023).

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{4}$$

O desvio padrão é simplesmente a raiz quadrada da variância e é representado pelo símbolo σ. Entretanto, saber puramente o valor do desvio padrão não permite saber se essa variabilidade é alta ou não nesse conjunto de dados específico. Assim, essa medida é mais facilmente interpretável se comparada com a média, surgindo então o coeficiente de variação (CV), que é calculado pelo percentual do desvio padrão em relação à média (5) (DOS SANTOS, 2023).

$$CV = \frac{\sigma}{\bar{x}}100 \tag{5}$$

Vale mencionar que o IQR também fornece uma perspectiva sobre a variabilidade dos dados.

Todas as medidas exibidas acima focavam na análise de uma única variável, sem avaliar sua interação com demais variáveis do conjunto. Para isso, introduz-se o conceito de correlação, que consiste de uma medida estatística usada para quantificar a força e a direção da relação entre duas variáveis.

A correlação de Pearson é a mais comumente utilizada e seu coeficiente mede o grau de correlação linear entre duas variáveis quantitativas *x* e *y* (GRAVETTER e WALLNAU, 2017). Seu valor pode variar de -1 a 1 e é calculado através da fórmula abaixo:

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{6}$$

Onde x_i e y_i são os valores das variáveis e $\bar{x}$ e $\bar{y}$ são as médias de cada uma. Quanto mais próximo de 1, maior a correlação positiva e quanto mais próximo de -1, maior a correlação negativa. Uma correlação próxima de zero significa uma relação linear entre as duas variáveis quase inexistente.

Em contrapartida, a correlação de Spearman é uma medida não paramétrica de correlação de ranque (GRAVETTER e WALLNAU, 2017). Seu coeficiente é usado para medir a força e direção da associação monotônica entre duas variáveis, que pode ou não ser linear. Seu coeficiente é obtido pela fórmula abaixo e a interpretação dos valores é análoga à do coeficiente de correlação de Pearson.

$$r_s = \frac{6 \sum d_i}{n(n^2 - 1)} \tag{7}$$

Na equação acima, d_i é a diferença entre os ranques das variáveis em cada observação e n é o número de observações.

A correlação de Pearson é mais apropriada quando as variáveis são aproximadamente normalmente distribuídas e a relação entre elas é linear, enquanto Spearman é mais robusta em relação a outliers e pode ser usada para quando a relação entre variáveis não é linear. Ambas as correlações são ferramentas poderosas na análise de dados, permitindo identificar e quantificar a força das relações entre variáveis.

Por fim, é necessário discorrer minimamente sobre a normalização de variáveis. A normalização de variáveis, particularmente a normalização pelo z-score, é uma técnica estatística usada para padronizar dados. O z-score (também conhecido como padronização) transforma os dados para que tenham uma média de zero e um desvio padrão de um. Este processo é crucial em muitas análises estatísticas, especialmente em métodos que assumem que os dados estão distribuídos normalmente ou que requerem que todas as variáveis tenham a mesma escala (GRAVETTER e WALLNAU, 2017).

Mantendo as notações de média e desvio padrão utilizadas anteriormente, podemos definir a normalização de uma variável *X* através do z-score pela equação abaixo:

$$z = \frac{X - \bar{x}}{\sigma} \tag{8}$$

Após a padronização, cada observação reflete quantos desvios padrão seu valor original está distante da média do conjunto de dados. Um z-score de 0 indica que o valor da observação é igual à média, um z-score positivo indica um valor acima da média, e um z-score negativo indica um valor abaixo da média.

A normalização pelo z-score é particularmente útil em análises multivariadas, como a regressão linear múltipla e a análise de componentes principais, onde é importante comparar variáveis que podem estar em diferentes escalas ou ter diferentes unidades de medida.

b. Séries temporais

A análise de séries temporais é um componente fundamental da ciência de dados e do aprendizado de máquina, especialmente quando se trata de dados indexados e ordenados no tempo. Em séries temporais, as observações são realizadas em intervalos de tempo regulares ou irregulares, e a ordem temporal das observações é crucial para a análise e modelagem (HYNDMAN e ATHANASAPOULOS, 2018).

Ou seja, uma série temporal pode ser constituída de qualquer conjunto de dados onde a dimensão temporal é essencial, como preços de ações ao longo do tempo, dados meteorológicos, ou evolução de casos de uma doença (veja Figura 3). A análise de séries temporais envolve técnicas específicas para lidar com a dependência temporal, tendências, sazonalidade e outros aspectos dinâmicos dos dados.

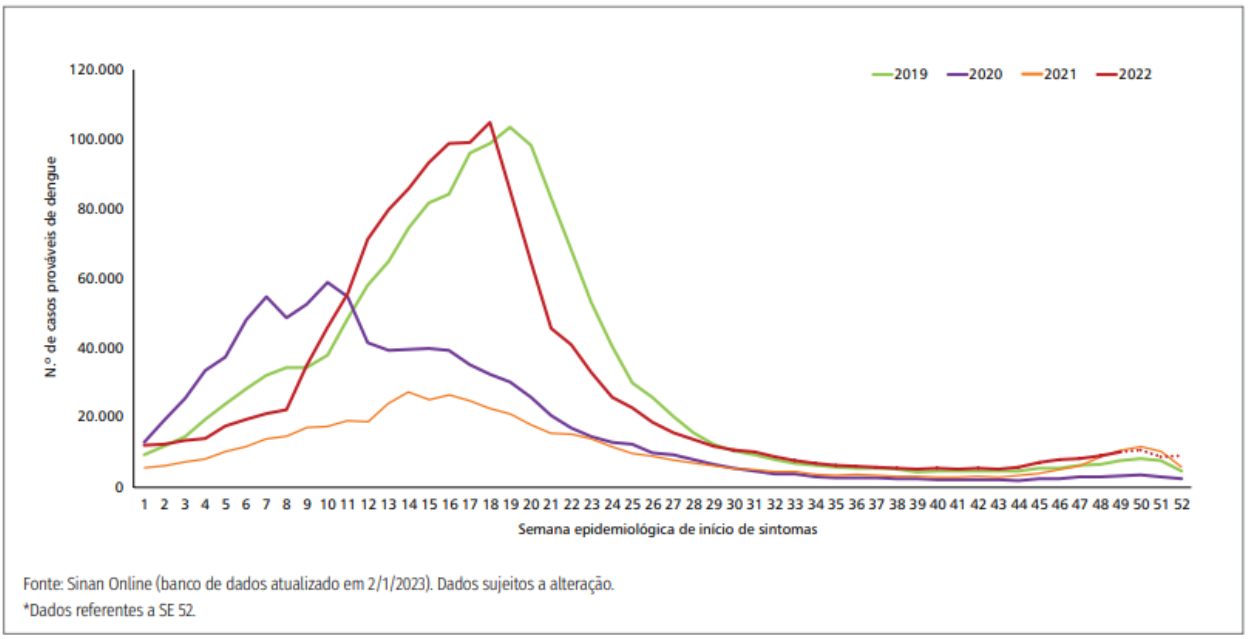

Figura 3- Série temporal de casos prováveis de dengue
Fonte: Secretaria de Vigilância em Saúde e Ambiente; Ministério da Saúde, 2023

Nesse contexto, a definição da estacionariedade é crucial, pois muitos métodos de análise e previsão de séries temporais assumem que a série é estacionária, ou pelo menos aproximadamente estacionária. Uma série temporal estacionária é aquela cujas propriedades estatísticas, como média, variância e autocorrelação, permanecem constantes ao longo do tempo (HYNDMAN e ATHANASAPOULOS, 2018).

Uma série é dita fracamente estacionária se sua média e variância são constantes ao longo do tempo e a covariância entre dois períodos de tempo depende apenas da distância (ou lag) entre esses períodos, e não dos tempos específicos em que as observações foram feitas.

Uma série é estritamente estacionária se as propriedades estatísticas de qualquer conjunto de períodos de tempo são as mesmas que as de qualquer outro conjunto idêntico de períodos, deslocado no tempo. Isso significa que todas as propriedades da série (distribuições de probabilidade) são invariantes no tempo.

Muitas séries temporais reais não são estacionárias, mas podem ser transformadas em séries estacionárias por meio de métodos como a diferenciação, a aplicação de transformações logarítmicas ou a remoção de componentes de tendência e sazonalidade.

Para isso, é necessário entender o modelo de decomposição clássico, no qual uma série é transformada para um formato mais adequado de ser trabalhado (HYNDMAN e ATHANASAPOULOS, 2018). Considerando uma série temporal $\{X_k\}_{k\in N}$, presume-se que essa série pode ser decomposta da seguinte forma:

$$X_k = m_k + s_k + Y_k \tag{9}$$

Na equação acima, Y_k é uma série estacionária, s_k é o componente sazonal e m_k é o componente de tendência. A sazonalidade refere-se a flutuações que ocorrem em intervalos regulares devido a fatores sazonais – ou seja, é uma componente periódica no tempo. Já a tendência representa a componente de longo prazo da série temporal, mostrando um padrão persistente ao longo do tempo. A tendência pode ser linear ou não linear e reflete o aumento ou diminuição dos valores da série ao longo do tempo de forma não periódica (HYNDMAN e ATHANASAPOULOS, 2018).

Para estimar a sazonalidade e, algumas técnicas podem ser usadas, como filtros de médias móveis e outros filtros passa-baixa, suavização exponencial e ajustes polinomiais. Após estimada, a tendência pode ser eliminada através de diferenciação e a sazonalidade através da média móvel.

AR (Autoregressive), MA (Moving Average) e ARMA (Autoregressive Moving Average) são modelos fundamentais na análise de séries temporais, especialmente úteis para modelar dados estacionários. Esses modelos são amplamente usados para análise e previsão de séries temporais em várias áreas, como economia, finanças, engenharia, entre outras (HYNDMAN e ATHANASAPOULOS, 2018).

Um modelo AR de ordem p, denotado como AR(p), usa a dependência entre uma observação e um número de atrasos (lags) dessa observação. O modelo é definido como:

$$Y_k = c + \beta_1 Y_{k-1} + \beta_2 Y_{k-2} + \cdots + \beta_p Y_{k-p} + \varepsilon_t \tag{10}$$

Um modelo MA de ordem q, denotado como MA(q), modela a série temporal como uma média móvel de termos de erro passados. A equação de um MA(q) é:

$$Y_k = c + \varepsilon_k + \emptyset_1 \varepsilon_{k-1} + \emptyset_2 \varepsilon_{k-2} + \cdots + \emptyset_q \varepsilon_{k-q} \tag{11}$$

O modelo ARMA combina ambos, AR(p) e MA(q), e é denotado como ARMA(p, q). O modelo ARMA pode capturar a autocorrelação em séries temporais usando componentes autoregressivos e de média móvel. A equação do modelo ARMA(p, q) é:

$$Y_k = c + \varepsilon_k + + \beta_1 Y_{k-1} + \beta_2 Y_{k-2} + \cdots + \beta_p Y_{k-p} + \emptyset_1 \varepsilon_{k-1} + \emptyset_2 \varepsilon_{k-2} + \cdots + \emptyset_q \varepsilon_{k-q} \tag{12}$$

c. Ciência de dados

A Ciência de Dados, como campo multidisciplinar, está no centro da revolução analítica, transformando a maneira como compreendemos e utilizamos os dados. A Ciência de Dados combina elementos da mineração de dados, estatística, aprendizado de máquina e análise preditiva, aplicando-os a conjuntos de dados de grandes dimensões para extrair conhecimentos úteis e padrões ocultos (LESKOVEC, RAJARAMAN e ULLMAN, 2014). Esta abordagem é fundamental em uma era onde o volume de dados gerados excede em muito a capacidade humana de analisá-los manualmente. Tal ciência não é apenas uma ferramenta para entender grandes volumes de dados, mas também uma técnica para resolver problemas complexos e tomar decisões informadas em diversos setores, desde o comércio até a saúde pública.

Em outra perspectiva, a Ciência de Dados é descrita como uma disciplina emergente que se baseia em métodos quantitativos e em uma abordagem pragmática para resolver problemas reais (O'NEIL e SCHUTT, 2013). A sua essência está na aplicação prática de técnicas matemáticas e computacionais para resolver problemas específicos, enfatizando a importância da comunicação e da interpretação dos resultados. O campo da Ciência de Dados é, portanto, não apenas uma questão de manipulação de dados, mas também de entender e contextualizar os dados dentro de um quadro mais amplo, combinando habilidades técnicas com pensamento crítico e visão estratégica.

d. Aprendizado de máquina

O Aprendizado de Máquina (Machine Learning, ML), um campo dentro da Ciência de Dados, envolve o desenvolvimento de algoritmos que permitem que máquinas aprendam e façam previsões ou tomem decisões baseadas em dados. Ao contrário dos programas tradicionais, que executam instruções exatas, os algoritmos de ML se adaptam e melhoram com a experiência (BISHOP, 2006) que só pode ser alcançado através da análise de grandes volumes de dados e a identificação de padrões dentro destes.

No funcionamento do ML, dois aspectos predominam: abordagens algébricas e probabilísticas. As abordagens algébricas, como as Regressão Linear, utilizam funções matemáticas para processar e aprender a partir dos dados. Por outro lado, métodos probabilísticos, como os algoritmos de Bayes, lidam com a incerteza e a probabilidade para fazer previsões ou tomar decisões (BISHOP, 2006). Ambas as abordagens dependem fortemente de técnicas estatísticas e matemáticas para modelar e interpretar dados complexos.

Uma distinção essencial no ML é entre aprendizado supervisionado e não supervisionado. No aprendizado supervisionado, os algoritmos são treinados em um conjunto de dados rotulado, aprendendo a associar entradas específicas a saídas desejadas. Exemplos clássicos incluem modelos de regressão e classificação, usados em casos como de classificação de e-mails em "spam" ou "não spam" e previsão de preços de casas com base em características como tamanho e localização. Já no aprendizado não supervisionado, os algoritmos trabalham com dados não rotulados, identificando padrões e estruturas por conta própria (BISHOP, 2006). Um exemplo clássico é a segmentação de clientes em um mercado com base em comportamentos de compra, sem rótulos prévios.

Dentro do aprendizado de máquina supervisionado, dois dos principais modelos são a regressão e a classificação, que, embora compartilhem a característica de serem treinados com dados rotulados, diferem significativamente em suas aplicações e saídas. Modelos de regressão são usados quando o objetivo é prever valores contínuos ou quantitativos. Por outro lado, os modelos de classificação são empregados para prever categorias discretas. Aqui, o resultado é uma classificação em categorias pré-definidas. Enquanto a regressão estabelece uma relação funcional entre variáveis independentes e dependentes para prever um valor contínuo, a classificação categoriza os dados de entrada em classes distintas, com base nos padrões aprendidos durante o treinamento (BISHOP, 2006).

No contexto da dengue, modelos de regressão podem ser usados para prever o número de casos em uma cidade, enquanto modelos de classificação podem ajudar a predizer a probabilidade um indivíduo estar infectado ou não.

i. Principais algoritmos

Dentro da classe de algoritmos supervisionados de regressão, podemos citar alguns dos mais utilizados e que foram testados no trabalho em questão. Eles são: Regressão Linear, Árvore de Decisão, Random Forest e Gradient Boost.

A regressão linear é um dos métodos mais fundamentais e amplamente utilizados em estatística e aprendizado de máquina para prever um valor quantitativo (BISHOP, 2006). Baseia-se na premissa de que existe uma relação linear entre uma variável dependente (Y) e uma ou mais variáveis independentes (X).

Em sua forma mais simples, a regressão linear pode ser representada pela equação matricial abaixo.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{13}$$

Nessa equação, Y é a variável dependente (ou resposta), X é uma matriz com as variáveis independentes, β_0 é o intercepto, β_1 é o vetor dos coeficientes das variáveis independentes e ϵ é o erro.

Através do modelo de regressão linear, β_0 e β_1 são determinados algebricamente de tal forma que a soma dos quadrados dos resíduos seja mínima. Para descrever esse processo de otimização realizada na obtenção dos valores dos coeficientes e do intercepto, vamos definir $\hat{y}_i$ como sendo o valor predito de Y correspondente ao ponto x_i. Isto é:

$$\hat{y}_i = \beta_0 + \beta_1 x_i \tag{14}$$

O erro de cada ponto previsão é justamente a diferença entre o valor predito e o valor real de y (veja Figura 4). Ou seja, $y_i - \hat{y}_i$. Tendo isso em vista, o modelo calcula os coeficientes de tal forma que a soma do quadrado dos erros seja mínima (BISHOP, 2006). Em outras palavras, o modelo busca valores de β_0 e β_1 que minimizem a equação abaixo (15).

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{15}$$

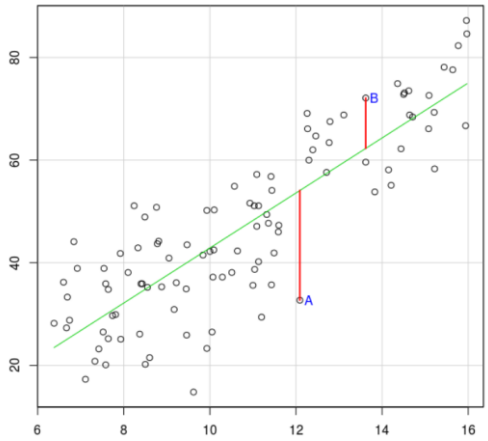

Figura 4 - Erro de previsão de um modelo
Fonte: FREIRE, 2021

Vale notar que, na construção do modelo de regressão linear, presume-se não só a linearidade dos dados, como também a normalidade, homoscedasticidade (variância constante) e independência dos erros. Logo, faz-se elementar validar essas pressuposições ao se aplicar um modelo de regressão linear.

Uma das grandes vantagens que a regressão linear confere é uma interpretação simples de seus resultados através dos coeficientes lineares das variáveis. Para ocasiões onde não só é necessário prever de forma precisa, mas explicar como cada fator influenciou na previsão, este algoritmo é amplamente recomendado.

Analogamente, outro algoritmo de entendimento facilitado para o usuário é a árvore de decisão. O algoritmo de Árvore de Decisão é uma técnica popular no campo do aprendizado de máquina, usada tanto para classificação quanto para regressão. Uma árvore de decisão é uma estrutura em forma de árvore que representa decisões e suas possíveis consequências na probabilidade do resultado (MITCHELL, 1997).

Uma árvore de decisão é composta de nós, ramificações e folhas. Cada nó interno representa um atributo ou uma pergunta, cada ramificação representa a decisão tomada, e cada folha representa um resultado ou uma classe (veja Figura 5). A árvore é construída dividindo o conjunto de dados em subconjuntos baseados em valores de atributos. Este processo é conhecido como "bifurcação" e é repetido de forma recursiva em cada subconjunto derivado.

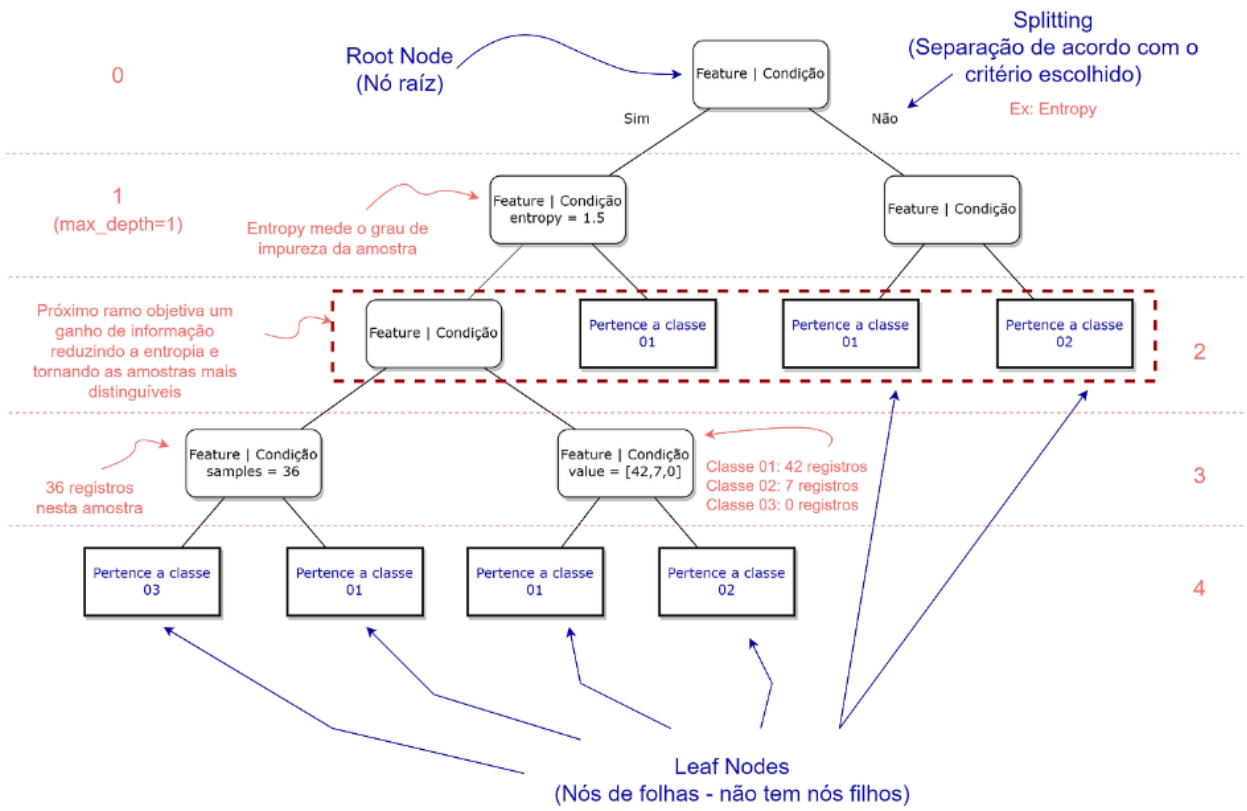

Figura 5 - Esquematização de árvore de decisão
Fonte: FREITAS, 2022

A seleção de atributos para a bifurcação é crucial e geralmente é feita com base em critérios como o índice de Gini e ganho de informação. Neste quesito, faz-se crucial explicar com mais detalhes em que consiste cada um desses critérios.

A começar pelo Índice de Gini, ele é usado para medir a impureza de um (sub)conjunto de dados. A pureza total significa que o subconjunto contém apenas dados de um único valor na variável dependente e é representada pelo índice 0, enquanto que é a pureza máxima é representada pelo índice 1 (ONODA e EBECKEN, 2001). Seja p_i a frequência relativa de determinada classe e c o número de classes, o índice de Gini é dado por:

$$Índice\ de\ Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{16}$$

Em paralelo, o ganho de informação é uma medida que remete à redução de entropia proporcionada por uma bifurcação, onde a entropia é definida pela fórmula (17). Considerando que a entropia representa a heterogeneidade da variável dependente, a redução desta significa uma contribuição para a previsão ao reduzir a aleatoriedade do resultado (MITCHELL, 1997).

$$Entropia(S) = -\sum_{i=1}^{c} p_i log_2(p_i) \tag{17}$$

Pela simplicidade de entendimento do algoritmo e das regras que regem as decisões, as árvores de decisão compartilham das mesmas vantagens da regressão linear no quesito simplicidade e explicabilidade, além de ser capaz de lidar com dados não lineares. No entanto, são propensas ao overfitting, especialmente quando são muito profundas. Para mitigar isso, técnicas como poda, limitação da profundidade da árvore ou mínimo número de amostras por folha são usadas (MITCHELL, 1997).

Em contraste com os algoritmos interpretáveis, o Random Forest é um algoritmo cujo resultado é de difícil interpretação devido à sua complexidade probabilística. Ele consiste em uma técnica de aprendizado

de máquina que combina múltiplas árvores de decisão para formar um modelo mais robusto e preciso. Para isso, a Random Forest opera pelo princípio do "wisdom of crowds", onde a agregação das previsões de várias árvores reduz o risco de overfitting presente em árvores individuais e aumenta a precisão do modelo (BREIMAN, 2001).

Em uma Random Forest, cada árvore é construída usando um subconjunto dos dados de treinamento, contendo uma seleção aleatória de variáveis independentes e uma amostra também aleatória dos dados (bootstrapping). Essa aleatoriedade ajuda a aumentar a diversidade entre as árvores na floresta, reduzindo a correlação entre elas e, consequentemente, diminuindo a variância do modelo sem aumentar significativamente o viés (BREIMAN, 2001).

A previsão de uma Random Forest é então feita tomando a média das previsões de todas as árvores individuais (veja Figura 6), dado pela equação (18).

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i \tag{18}$$

Onde $\hat{y}_i$ é a previsão da i-ésima árvore e N o número total de árvores usadas no modelo.

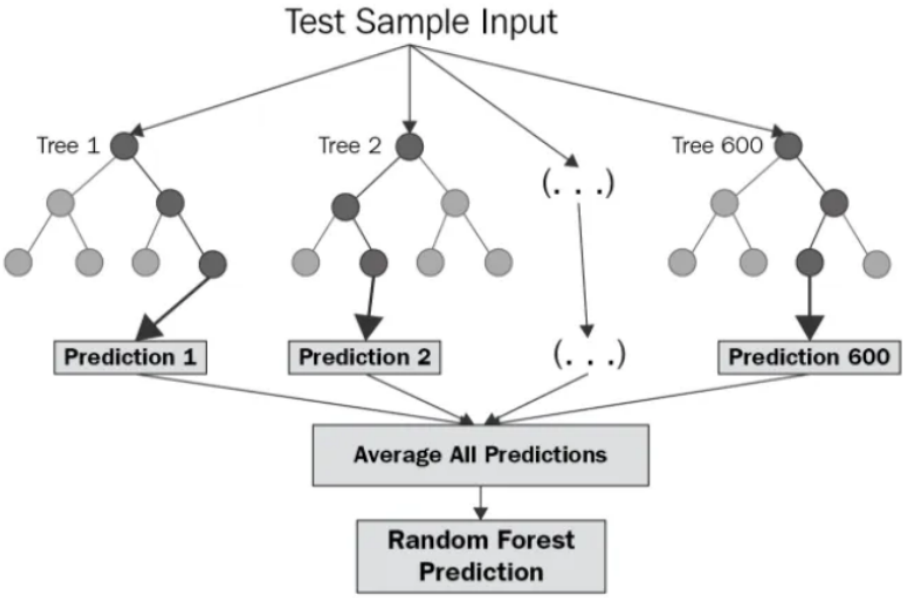

Figura 6 - Funcionamento do Random Forest
Fonte: Level up coding, 2020

Uma das principais vantagens da Random Forest é sua habilidade de lidar com grandes conjuntos de dados com dimensões variáveis (muitos recursos) e com dados que podem não seguir premissas de linearidade (BREIMAN, 2001). Por mais que a interpretação dos resultados seja complexa devido à quantidade de árvores e ao processo de bootstrap sampling, essa técnica pode fornecer medidas de importância das variáveis, que indicam quais variáveis são mais significativas na previsão.

Por fim, o Gradient Boosting é uma técnica elaborada através de um processo bastante similar ao Random Forest. Ele é um método de aprendizado de máquina que cria um modelo preditivo na forma de um conjunto de modelos mais fracos (geralmente árvores de decisão) em sequência, visando sempre corrigir o erro do modelo anterior (MURPHY, 2012).

O processo do Gradient Boosting começa com um modelo inicial, que pode ser tão simples quanto uma previsão constante. Em cada etapa subsequente, o algoritmo adiciona uma árvore ao modelo existente de forma que minimize a função de perda total. A função de perda avalia quão bem o modelo está se saindo com relação aos dados de treinamento. A ideia central é construir cada nova árvore para corrigir os erros cometidos pela sequência das árvores anteriores. Matematicamente, cada nova árvore é

construída para prever o gradiente negativo da função de perda com relação às previsões, daí o nome "Gradient Boosting". A previsão final é feita através da soma ponderada das previsões de todas as árvores (MURPHY, 2012).

Assim, por mais que tanto o Random Forest quanto o Gradient Boosting utilizem o resultado de diversas árvores de decisão, uma das principais diferenças se dá pela forma como é construído. Enquanto o Random Forest cria um grande número de árvores independentes treinadas em cima de amostras aleatórias, o Gradient Boosting se apoia no conceito de criar árvores sequencialmente, onde as árvores subsequentes aprendem a partir dos erros das árvores anteriores, corrigindo-os. A Figura 7 e uma esquematização do funcionamento do processo de Gradeint Boosting (MURPHY, 2012).

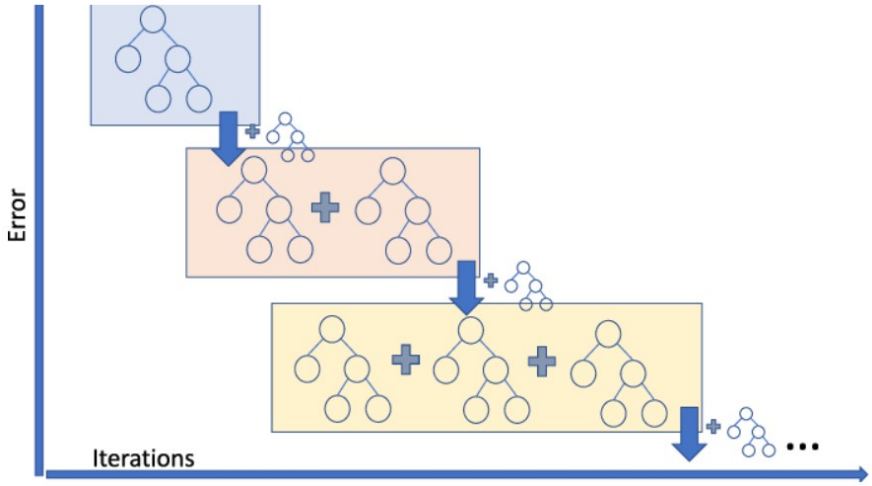

Figura 7 – Representação esquematizada do algoritmo de Gradient Boosting
Fonte: BATURYNSKA e MARTINSEN, 2021

O Extreme Gradient Boosting (XGBoost) é uma extensão do Gradient Boosting, diferenciada pela eficiência computacional, precisão, e a capacidade de ser distribuído em vários computadores, tratando de grandes conjuntos de dados e recursos de maneira mais eficaz (CHEN e GUESTRIN, 2016). O XGBoost introduz várias melhorias como regularização (L1 e L2), que ajuda a prevenir overfitting, e um procedimento mais eficiente para avaliar e dividir os nós das árvores, conhecido como "quantile sketch". Além disso, o XGBoost pode lidar automaticamente com valores faltantes e oferece suporte para várias funções de perda, o que o torna aplicável a uma ampla variedade de problemas.

ii. Overfitting e underfitting

Overfitting e underfitting são dois problemas comuns que podem ocorrer durante o treinamento de modelos de aprendizado de máquina, impactando negativamente sua capacidade de generalizar bem para novos dados.

Overfitting ocorre quando um modelo é treinado de forma excessiva com os dados de treinamento, aprendendo tanto os padrões legítimos quanto o ruído presente neles. Isso significa que o modelo se torna extremamente ajustado ou adaptado aos dados de treinamento e, como resultado, tem um desempenho ruim em dados novos, como os de teste. Overfitting é particularmente comum em modelos muito complexos, como redes neurais profundas, que têm um grande número de parâmetros e a capacidade de aprender representações muito detalhadas dos dados de treinamento. Um modelo overfitting pode ser identificado quando há uma grande discrepância entre o desempenho nos dados de treinamento e nos dados de teste - alto desempenho no treinamento, mas baixo desempenho no teste – e é comum em bases de dado com muitas variáveis preditivas se comparadas ao número de observações usadas para treinamento (MURPHY, 2012).

Underfitting, por outro lado, ocorre quando um modelo é muito simples para capturar a estrutura subjacente dos dados. Isso pode ser devido a uma escolha inadequada de modelo, falta de características relevantes nos dados, ou porque o modelo não foi treinado o suficiente. Como resultado, o modelo falha tanto no treinamento quanto nos dados de teste, indicando que ele não aprendeu suficientemente bem

os padrões nos dados de treinamento (DE SOUZA, 2022). A Figura 8 apresenta alguns exemplos desses conceitos.

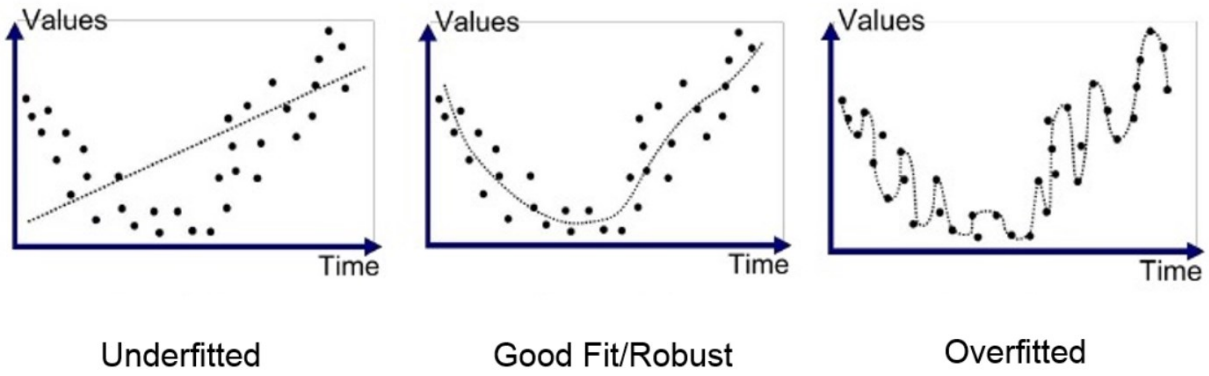

Figura 8 - Modelos overfitted e underfitted
Fonte: BRANCO, s.d.

Para evitar overfitting e underfitting, práticas como a validação cruzada, a seleção cuidadosa de características, a regularização (para limitar a complexidade do modelo), e o ajuste de parâmetros são cruciais. Além disso, é importante ter um equilíbrio entre a complexidade do modelo e a quantidade e variedade dos dados de treinamento.

iii. Métricas de avaliação do modelo

No aprendizado de máquina, especialmente em tarefas de regressão, é crucial avaliar a performance dos modelos usando métricas apropriadas. Três das métricas mais comuns são o Erro Médio Absoluto (MAE), o Erro Quadrático Médio (MSE) e o coeficiente de determinação R².

O MAE é a média do módulo dos erros entre as previsões e os valores reais (O'NEIL e SCHUTT, 2013). Seguindo a mesma notação anterior, onde $\hat{y}_i$ são as previsões e y_i são os valores reais, o MAE pode ser calculado segundo a fórmula abaixo:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{19}$$

Analogamente, o MSE é calculado através da média dos quadrados dos erros entre previsto e realizado (HASTIE, TIBSHIRANI e FRIEDMAN, 2016). Assim, o MSE é obtido através da fórmula abaixo:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{20}$$

Em contraste com o MAE, o MSE é mais sensível à outliers, pois erros maiores são mais penalizados nessa métrica devido ao fator quadrático envolvido. Em outras palavras, para reduzir o MSE, são preferíveis erros constantes e pequenos do que erros pontuais e grandes. Para ambas essas métricas, quanto maior, pior é a assertividade da previsão (HODSON, 2022).

Por fim, o R² mede a proporção da variância na variável resposta explicada pelas variáveis independentes. Um valor próximo de 1 significa que a modelo explica boa parte da variância do modelo, enquanto que um valor próximo de 0 indica o contrário. Mais precisamente, essa métrica compara a previsão do modelo com uma previsão constante feita através da média dos valores da variável dependente – que seria o modelo mais simples possível (HASTIE, TIBSHIRANI e FRIEDMAN, 2016).

A fórmula do R² pode ser vista abaixo, na qual $\hat{y}_i$ denota a média dos valores reais.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{21}$$

3. Desenvolvimento

O objetivo do trabalho, conforme descrito na introdução, foi desenvolver um modelo preditivo de dengue que atendesse o estado do Rio de Janeiro. A partir de estudo prévio acerca da doença e visando um modelo o mais completo possível, pretendia-se incluir variáveis socioeconômicas, climáticas, geográficas e, é claro, epidemiológicas para explicar o comportamento dos casos de dengue no estado.

Entretanto, como ponto de partida, foi necessário estruturar uma base de dados capaz de fornecer todas essas informações de forma confiável. A seção a seguir trata de como essa base foi construído e trabalhada, apresentando explicações sobre as fontes de dados utilizadas.

a. Fontes de dados

i. Caso de dengue no estado do RJ

A informação central deste estudo era quantidade de casos de dengue no Rio de Janeiro. Idealmente, para que fosse possível atingir os objetivos almejados com o trabalho, seria necessário encontrar uma fonte de dados históricos de casos de dengue em toda a unidade federativa do Rio de Janeiro.

Devido às limitações impostas por medidas de segurança da informação, era difícil obter dados granulares de casos de dengue com histórico suficiente. Nem mesmo dados na granularidade de bairros foram possíveis de serem encontrados, logo optou-se por trabalhar com informações a nível de município.

Além dessa questão, também era importante ter um histórico suficiente para se trabalhar, o que se estipulou como sendo de 5 anos. Ou seja, desde 2018.

Para que esse dado pudesse ser trabalhado com mais precisão, principalmente pensando no impacto de fatores climáticos e a variabilidade das condições meteorológicas no RJ, também era importante que a frequência dos registros fosse suficientemente alta, como de dia ou semana.

Tendo em vista os pré-requisitos acima, algumas possíveis fontes de dados foram listadas.

A primeira delas era o SINAN. O SINAN (Sistema de Informação de Agravos de Notificação) é um sistema brasileiro usado para a coleta e processamento de dados sobre doenças e agravos que são de notificação compulsória. Como a dengue é uma doença de notificação obrigatória, todas os registros de tal são unificados em uma base de dados orquestrado pelo SUS, denominada SINAN (Sistema de Notificação de agravos de notificação). Os dados sobre dengue do SINAN incluem informações detalhadas dos casos (suspeitos e confirmados), como dados demográficos do paciente, dados clínicos e laboratoriais do caso, informações epidemiológicas e temporais dos casos.

Ao que parecia, seria a base de dados ideal para o trabalho, não fosse o fato de que, em alguns anos, não havia quase nenhum preenchimento do campo de UF de notificação do caso. Além disso, os dados não estavam suficientemente atualizados.

Outras possíveis fontes de dados fornecidas pelo SUS era o SIA (Sistema de Informações Ambulatoriais) e SIH (Sistema de Informações Hospitalares). O SIA registra todos os procedimentos ambulatoriais realizados e o SIH as internações hospitalares, sendo possível filtrar o CID (Classificação Estatística Internacional de Doenças e Problemas Relacionados à Saúde).

Entretanto, havia alguns entraves em se trabalhar com essas bases. Podemos listar a baixa qualidade dos dados, atualizações defasadas e as etapas de pré-processamento necessárias para se chegar ao dado principal, que eram os casos de dengue. Além disso, no caso do SIH, só havia registros de casos graves que exigiram internação, o que não era o alvo do estudo.

Felizmente, muito em vista do grande enfoque e número de estudos feitos em cima da dengue, já havia algumas bases de dados tratadas com informações sobre casos de dengue. Nesse sentido, a Fiocruz (Fundação Oswaldo Cruz) desempenhou um papel crucial de mantenedora de bases de dados na área da saúde, fomentando pesquisas em cima das mesmas.

Uma dessas bases nasceu do InfoDengue. O InfoDengue é um sistema inovador de alerta para arboviroses desenvolvido pela FioCruz em parceria com a FGV e Observatório da Dengue (UFMG). Ele se baseia em dados de redes sociais, climáticos, demográficos e epidemiológicos, combinados com técnicas

de inteligência artificial e modelagem matemática para gerar previsões precisas sobre o risco de transmissão dessas doenças.

Além disso, o InfoDengue disponibiliza essas informações de forma acessível, facilitando a tomada de decisões baseada em evidências para combater essas doenças transmitidas por mosquitos. Os dados disponibilizados pelo InfoDengue estão sumarizados à nível de semana epidemiológica e município e incluem variáveis já pré-processadas.

Uma delas, que se mostrou importante para este trabalho, é a de quantidade de menções a sintomas de dengue, dentre tweets geolocalizados. Outra variável fornecida pelo sistema é o nível de receptividade do clima para transmissão de dengue, que leva em conta também por quanto tempo o clima tem se mantido receptivo, permitindo assim um ciclo completo de transmissão.

ii. Dados climáticos e meteorológicos

Conforme mencionado anteriormente, a transmissão da dengue é fortemente condicionada pelas condições climáticas, dentre as quais podemos citar prioritariamente a temperatura, umidade e precipitação. A base de dados do InfoDengue já fornecia variáveis de temperatura e umidade, mas ainda era necessário enriquecer com informações pluviométricas.

Para isso, primeiramente buscou-se dados do INMET (Instituo Nacional de Meteorologia). Entretanto, os dados fornecidos por essa fonte apresentaram dois principais problemas. O primeiro era a falta de padronização da estrutura dos dados ao longo dos anos e a baixa qualidade dos dados, o que exigia um trabalho moroso de sanitização dessa base. O outro problema era que o INMET só trabalha com dados à nível de estação meteorológica, que não tem uma relação de um para um com municípios, conforme podemos ver na Figura 9.

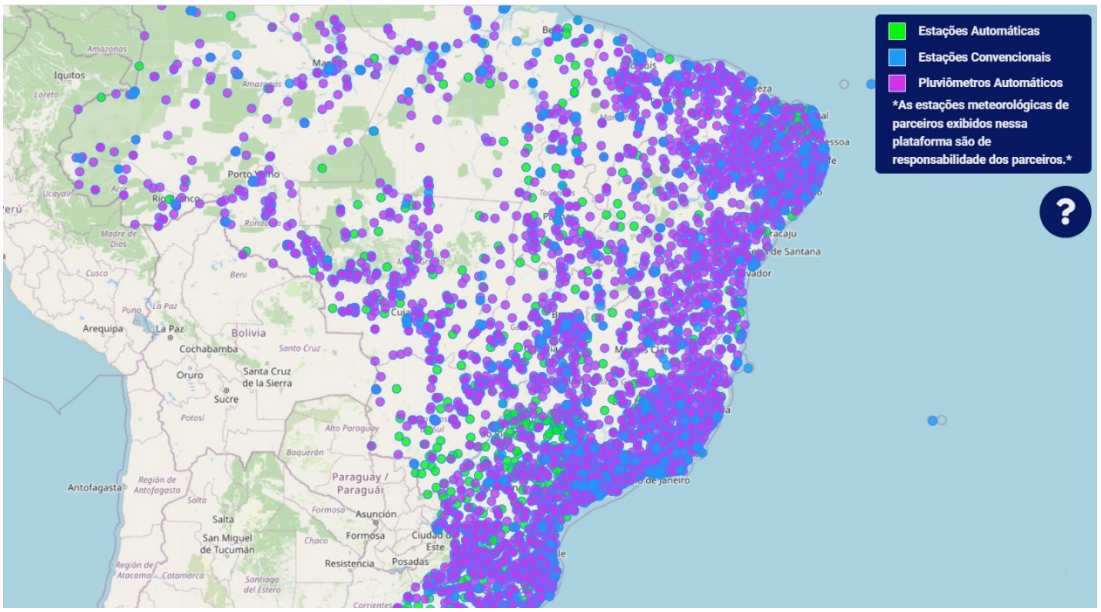

Figura 9 - Estações meteorológicas INMET
Fonte: INMET, 2023.

Desta forma, não havia estação em todo município fluminense, assim como em alguns municípios havia mais de uma estação.

Buscando uma alternativa mais direcionada, recorremos aos dados providos pelo OpenMeteo, plataforma que infere diversos indicadores meteorológicos por localização com base nos dados oficiais das estações. Assim, era possível atribuir condições climáticas diárias à cada município do RJ.

Os dados disponibilizados eram de boa qualidade e padronizados, mas em contrapartida sua exportação era manualmente custosa, pois cada arquivo gerado correspondia à um único município. Ao fim, a tabela gerada possuía informações diárias à nível de cidade com variáveis de temperatura (real e aparente) mínima, média e máxima e precipitação total.

iii. Dados socioeconômicos

Havia também a necessidade de enriquecer a base de dados com informações socioeconômicas dos municípios, as quais poderiam ser usadas para diferenciar quantitativamente cada local. Além disso, estudos indicam (SOUSA et al., 2016) que tais fatores desempenham um papel central nos padrões de contaminação por dengue, pois indicadores como grau de urbanização, densidade populacional, IDH e condições de saneamento básico impactam diretamente no nível de casos de determinado local.

Com esse intuito, foram utilizadas duas fontes de dados socioeconômicos. As informações de IDH foram colhidas do Atlas Do Desenvolvimento Humano no Brasil, que consolida índices de educação, renda, população, IDH-M, saúde e outros desde 1991. A nível municipal, são aproximadamente 120 indicadores disponíveis que ajudam a avaliar o grau de desenvolvimento humano.

O Índice de Desenvolvimento Humano Municipal (IDHM) é uma medida composta de indicadores de três dimensões do desenvolvimento humano: longevidade, educação e renda. O índice varia de 0 a 1. Quanto mais próximo de 1, maior o desenvolvimento humano (UNDP,2023).

Além dela, ainda foram utilizados dados do censo de 2010, que contribuíram com variáveis como população total dos municípios, percentual de população urbana e densidade populacional.

Juntando ambas, foi possível obter uma visão socioeconômica e demográfica bem ampla acerca dos 92 municípios que compõe o estado do Rio de Janeiro.

b. Preparação da base

As bases originalmente obtidas continham muitas sujeiras e exigiam um processo de sanitização para que atingissem a qualidade adequada para serem utilizadas em um modelo preditivo. Além disso, muitas vezes foi necessário criar novas variáveis a partir daquelas já prontas e que fossem capazes de ajudar a explicar a curva de casos de dengue com maior assertividade.

Assim, boa parte do desenvolvimento do projeto, como é de esperar em qualquer produto de dados, consistiu em uma preparação da base de dados, a qual pode ser decomposta em duas partes: limpeza de dados e engenharia de variáveis.

Cronologicamente falando, ambas as fases se intercalam, acompanhadas por constantes análises descritivas dos dados. Conforme os dados são evidenciados, é possível notar impurezas e entender comportamentos das variáveis que permitem transformar a base em uma mais adequada para o propósito desse trabalho.

i. Exploração e tratamento dos dados

A começar pelos dados do InfoDengue, a base utilizada apresentava dados de 2018 à novembro de 2023 de todos os 92 municípios do Rio de Janeiro (veja Figura 10 abaixo).

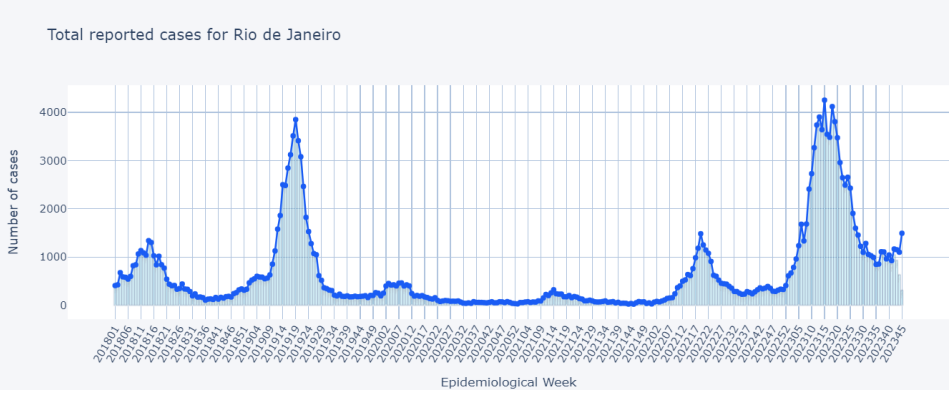

Figura 10 - Evolução de casos de dengue no Rio de Janeiro de 2019 à 2023

Conforme podemos ver no gráfico acima, o número de casos no estado do Rio de Janeiro apresentou um grande salto em 2023 se comparado aos anos anteriores. O único ano, dentre os compreendidos pelo estudo, com comportamento similar foi o de 2019.

A base de dados do Infodengue já fornecia algumas variáveis bem interessantes. Uma dela é a quantidade de tweets mencionando sintomas de dengue. Como podemos ver na Figura 11, a curva de tweets não acompanha diretamente a quantidade de casos, porém chama a atenção como o número de tweets foi representativo em 2019, crescendo junto com o número de casos.

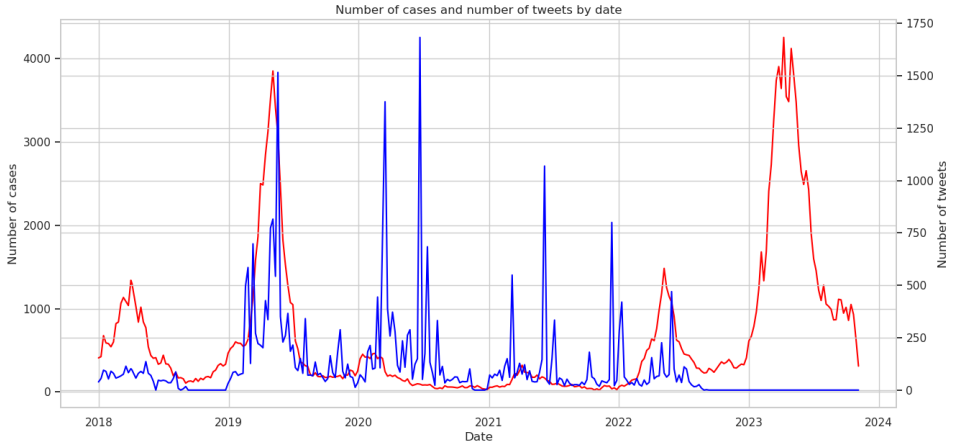

Figura 11 - Número de tweets vs. casos de dengue no Rio de Janeiro

Tal relação entre as variáveis foi comprovada no gráfico de dispersão da Figura 12. Nele, podemos notar que existe alguma correlação entre o número de tweets e a quantidade de casos, sendo o coeficiente de correlação de Pearson igual à 0,33. Por mais que não seja um alta correlação, isso mostra que essa variável deve ser levada em consideração para explicar o número de casos.

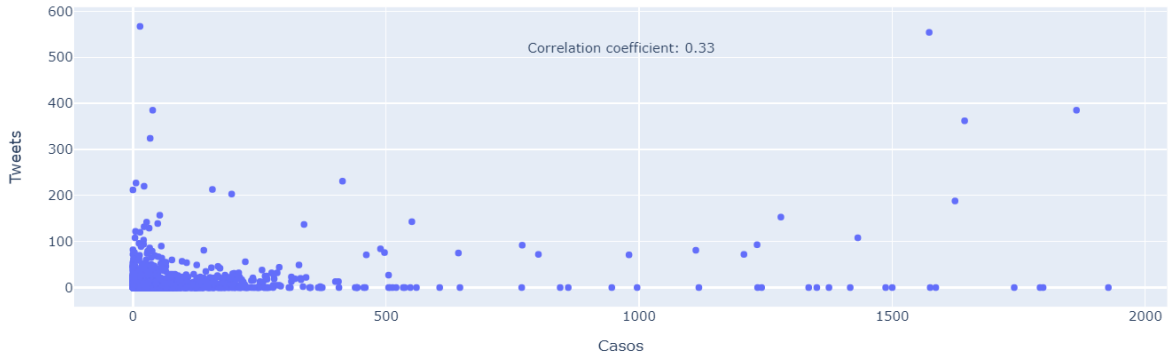

Figura 12 - Dispersão entre casos de dengue e tweets

Obtendo uma estatística descritiva básica de algumas variáveis numéricas, chama a atenção alguns fatos. Primeiro, quando olhamos para o número de casos, vemos que, apesar da média de casos ser próxima de 7 casos por semana e município, as 3º quartil é 0, o que mostra como essa variável é inflada de zeros. No decorrer do trabalho, foi importante levar isso em consideração.

Além disso, encontramos muitas colunas na base de dados com número expressivo de dados faltantes. Essas colunas foram eliminadas da base.

A próxima variável analisada foi a receptividade do clima para transmissão de dengue, calculada pelo sistema do Infodengue através das condições climáticas nas semanas anteriores. O gráfico da Figura 13

nos leva a acreditar que, quando o clima é receptivo, o média de casos consideravelmente maior do que no caso contrário.

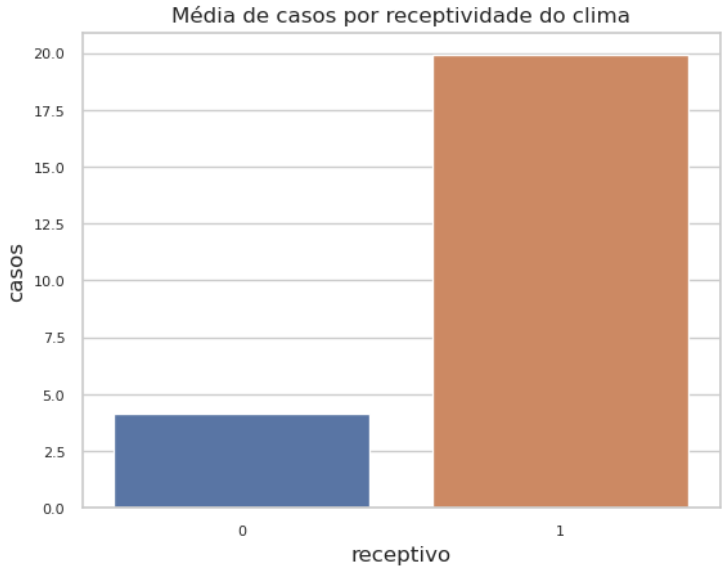

Figura 13 - Média de casos de dengue por semana epidemiológica de acordo com receptividade do clima

A fim de testar estatisticamente essa hipótese, primeiro checamos, através de um teste de Shapiro, se o número de casos era normalmente distribuído dentro dos dois grupos, o que não se mostrou verdadeiro. Em virtude disso, foi performado um teste não paramétrico de Mann-Whitney, que confirmou a hipótese inicial.

Em contrapartida, ao se analisar isoladamente as variáveis climáticas contidas nesse dataset, não foi possível estabelecer a conclusão de que alguma delas fosse impactante diretamente na transmissão de dengue. Para não enviesar a análise segundo dados de municípios mais populosos, focamos em analisar tais variáveis em comparação com a incidência de casos por 100 mil habitantes. As Figuras 14 e 15 mostram a desprezível correlação entre umidade e temperatura média em função da incidência.

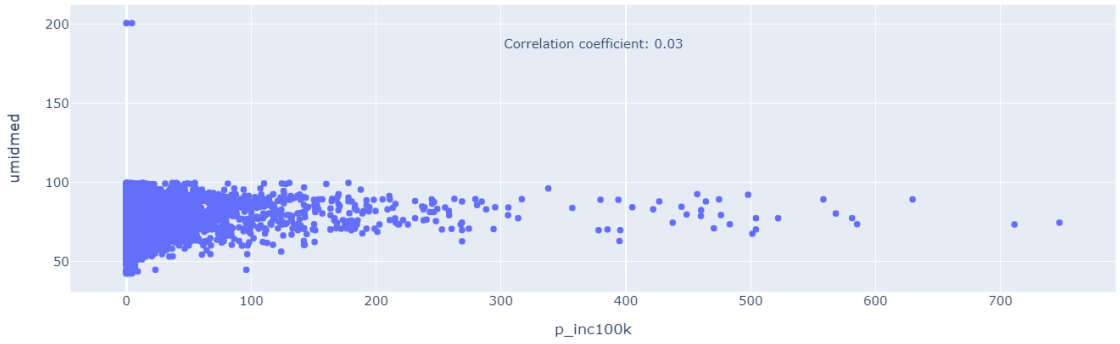

Figura 14 - Correlação entre incidência e umidade média

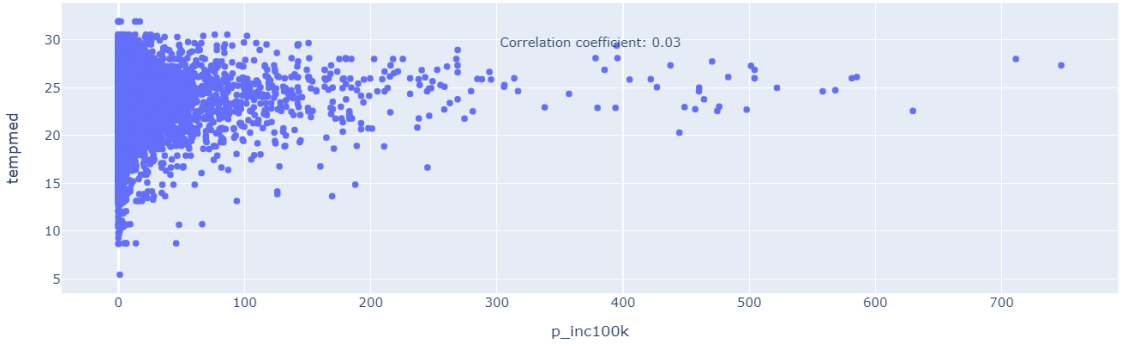

Figura 15 - Correlação entre incidência e temperatura média

Ou seja, é relevante considerar a continuidade de condições favoráveis para transmissão, mas não elas isoladamente na data da notificação, pois devemos levar em conta o ciclo de desenvolvimento e atividade dos mosquitos Aedes Aegypti e defasagem entre contaminação e notificação. De qualquer forma, é possível notar que as semanas de maior incidência de casos ocorreram em semanas com temperaturas entre 20 e 30º Celsius.

Apesar disso, temperatura e umidade não são os dois únicos fatores climáticos que influencia na transmissão de dengue. Os mosquitos Aedes Aegypti depositam suas larvas em focos de água parada e, portanto, é necessário água para que a quantidade de vetores da dengue aumente e a contaminação também consequentemente.

Sendo assim, foram utilizados dados meteorológicos do OpenMeteo para enriquecer a base de dados com informações pluviométricas. Essa base de dados fornecia informações detalhadas de temperatura, umidade e precipitação diárias nos 92 municípios do RJ de janeiro de 2018 à junho de 2023.

Podemos notar que a temperatura varia bastante de acordo com o município no boxplot da Figura 16. Esmo assim, a maioria das cidades apresenta uma mediana entre 20 e 25º, com variações entre os demais quartis.

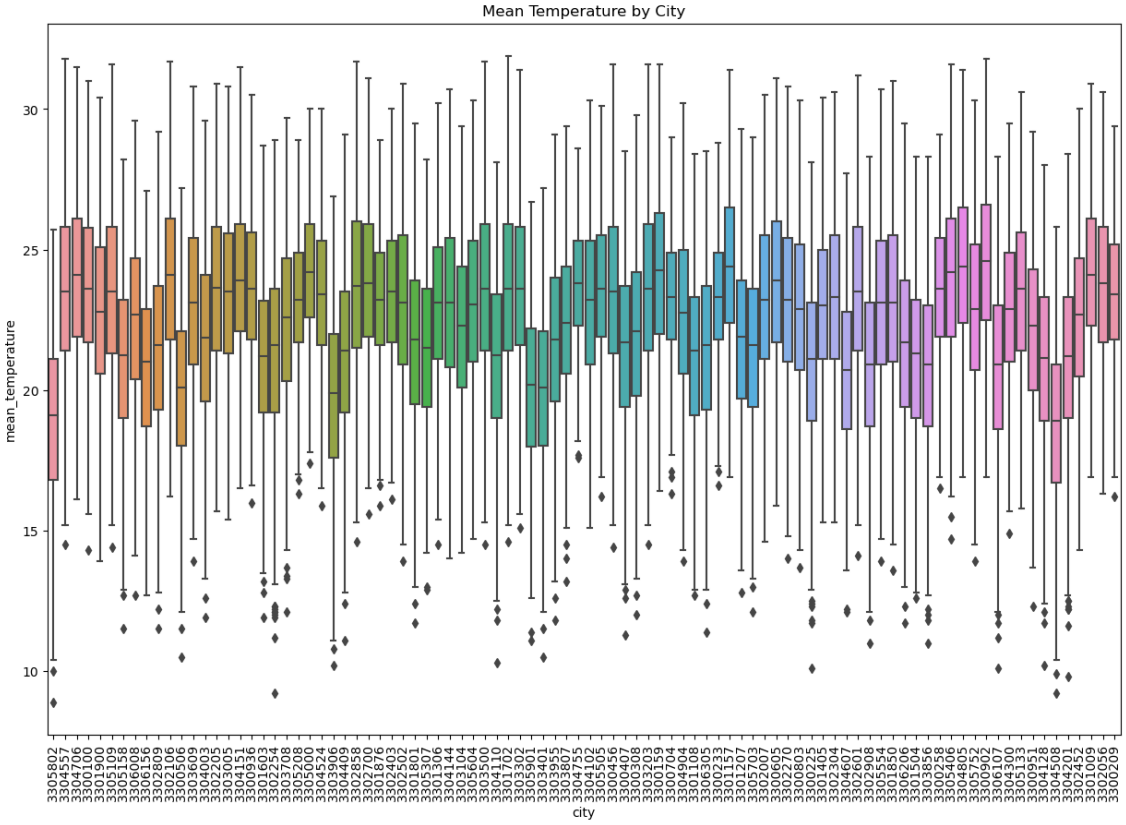

Figura 16 - Boxplot de temperatura média semanal por município

Para facilitar a análise das variáveis climáticas temporalmente, foram aplicadas médias móveis de 7 dias, o que facilitaria futuramente também no relacionamento desses dados com a base do InfoDengue. No gráfico da Figura 17, podemos claramente notar um comportamento sazonal da temperatura e da precipitação ao longo do ano, o que desperta uma possibilidade de isto se relacionar com o comportamento sazonal da curva de casos de dengue.

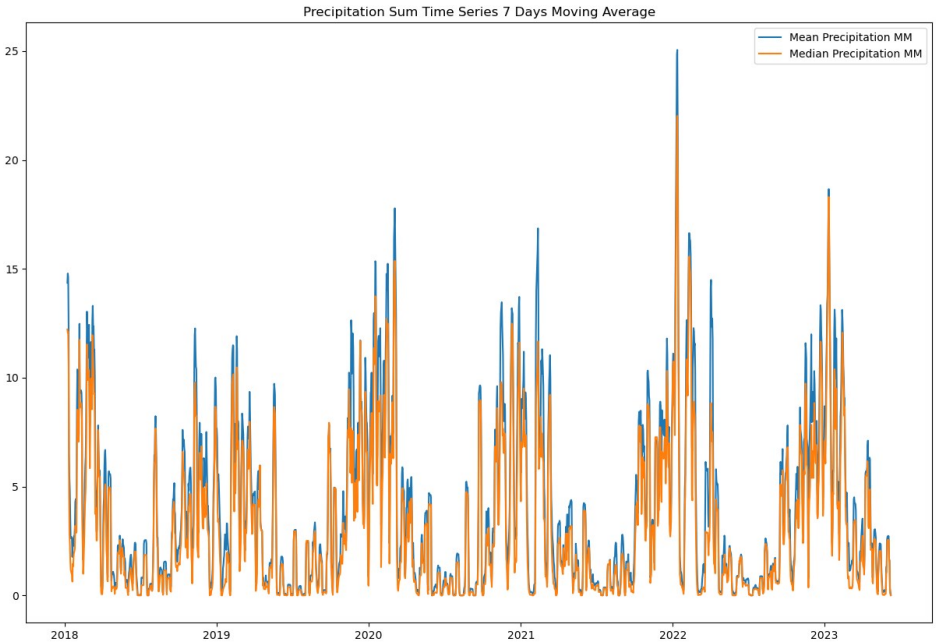

Figura 17 - Soma móvel de 7 dias de precipitação no RJ

Feito isso, a base do InfoDengue foi enriquecida com esses dados de acordo com a data de início da semana epidemiológica e o município. Logo, as variáveis acrescentadas se referiam às médias daquela variável climática ao longo da semana epidemiológica e no município em questão.

Na sequência, foram adicionadas as variáveis socioeconômicas vindas da base do Atlas de Desenvolvimento Humano. Há de se mencionar que essa base dispunha de um número muito elevado de colunas (1066), muitas das quais apresentavam um número demasiadamente alto de valores nulos. Mais especificamente, 996 dessas colunas eram compostas por mais de 99% de valores vazios. Logo, essas colunas foram eliminadas da base, restando 70 colunas para se trabalhar.

Entendendo com mais detalhes o que cada uma dessas 70 colunas representava, optou-se por trabalhar com um número reduzido de variáveis, as quais seriam analisadas com mais rigor. Essas variáveis estão descritas na Tabela 1.

Variável	**Descrição**
IDHM	Índice de Desenvolvimento Humano Municipal de 2010
IDHM_Renda	Dimensão de renda do IDH-M
IDHM_Educacao	Dimensão de educação do IDH-M
PIB_per_capita	Produto interno bruto per capita, segundo cálculos de 2016
Mortalidade DNT	Taxa de mortalidade por doenças não transmissíveis 2017
Internacoes_Saneamento	% de internações por doenças relacionadas ao saneamento ambiental inadequado 2017
IDHMAD	IDHM Ajustado à Desigualdade 2021
Domicilios_Saneamento	% de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados 2000

Tabela 1 - Descrição das variáveis sociais

Cada uma dessas variáveis foi analisada em comparação com a o número de casos e/ou a incidência, a fim de averiguar possíveis relações entre elas. Para isso, foram calculadas as correlações de Spearman, cujo resultado permite entender se existe alguma correlação entre os postos das variáveis, o que é interessante para casos em que as variáveis não seguem uma relação linear entre si. Além disso, também foram exibidos gráficos de dispersão destas variáveis com a incidência média por município.

Na Tabela 2 abaixo, podem ser vistos os resultados.

Variável	**Correlação de Spearman** (vs. p_inc100k)
IDHMAD	0,014
IDHM	0,194
IDHM_Renda	0,220
IDHM_Educacao	0,230
PIB_per_capita	0,111
Mortalidade DNT	-0,056
Internacoes_Saneamento	-0,006
Domicilios_Saneamento	0,051

Tabela 2 - Correlação de Spearman das variáveis sociais com a incidência média municipal

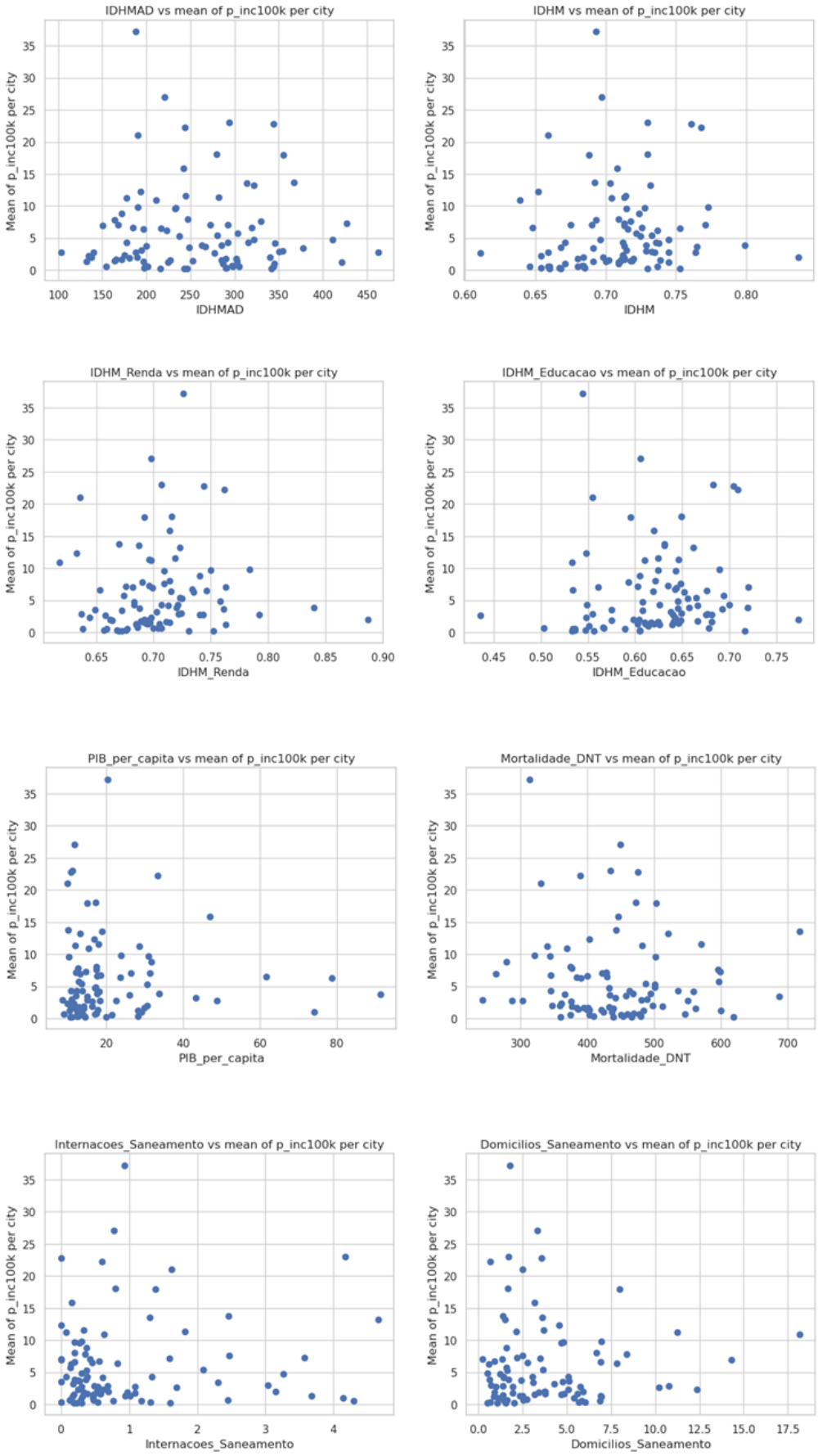

Figura 18 - Dispersão entre variáveis sociais e incidência média municipal

Tendo em vista os resultados acima, podemos notar que algumas variáveis tem baixíssima correlação de postos com a incidência de dengue no local. Por essa razão, decidimos desconsiderar do modelo variáveis com correlação inferior à 10%, o que faria com que restassem apenas os campos IDHM, IDHM_Renda, IDHM_Educacao e PIB_per_capita.

Entretanto, o que se notou é que algumas dessas variáveis eram altamente correlacionadas entre si, como podemos ver na matriz de correlação de Pearson da Figura 19. No sentido de utilizá-las conjuntamente em um modelo de regressão, isso não seria adequado devido à multicolinearidade. Ou seja, fatores redundantes acabariam por ser empregados no modelo, aumentando a complexidade e custo computacional sem aumentar proporcionalmente a capacidade preditiva.

	IDHM	IDHM Renda	IDHM Educacao	PIB per capita
IDHM	1	0,9	0,94	0,24
IDHM Renda	0,9	1	0,71	0,3
IDHM Educacao	0,94	0,71	1	0,15
PIB per capita	0,24	0,3	0,15	1

Figura 19 - Matriz de correlação das variáveis sociais selecionadas

É possível notar que IDHM, como era de se esperar, está fortemente correlacionado com os outros dois subindicadores de desenvolvimento humano, IDHM-Renda e IDHM-Educação. Entretanto, a correlação direta entre as duas últimas não é criticamente alta, assim como a variável PIB per capita não é altamente correlacionada com nenhuma das demais.

Desta forma, optou-se por remover a variável IDHM do modelo, mantendo apenas as outras três.

Em seguida, adicionamos à base de dados as informações populacionais extraídas do Censo de 2010 do IBGE. Dentre elas, com base em conhecimento acerca da doença, as mais relevantes eram população total, população urbana, percentual de população urbana e densidade populacional, cujas colunas respectivamente denominadas total, urbana, perc_urbana e densidade_pop.

Repetiu-se as análises feitas em cima das variáveis socioeconômicas para avaliar a importância desses campos para um futuro modelo de regressão, conforme podem ser vistas na Tabela 3 e Figuras 20 e 21.

Variável	**Correlação de Spearman** (vs. p_inc100k)
População total	-0,124
População urbana	-0,109
Percentual de população urbana	-0,054
Densidade da população	-0,103

Tabela 3 - Correlação das variáveis demográficas com a incidência média municipal

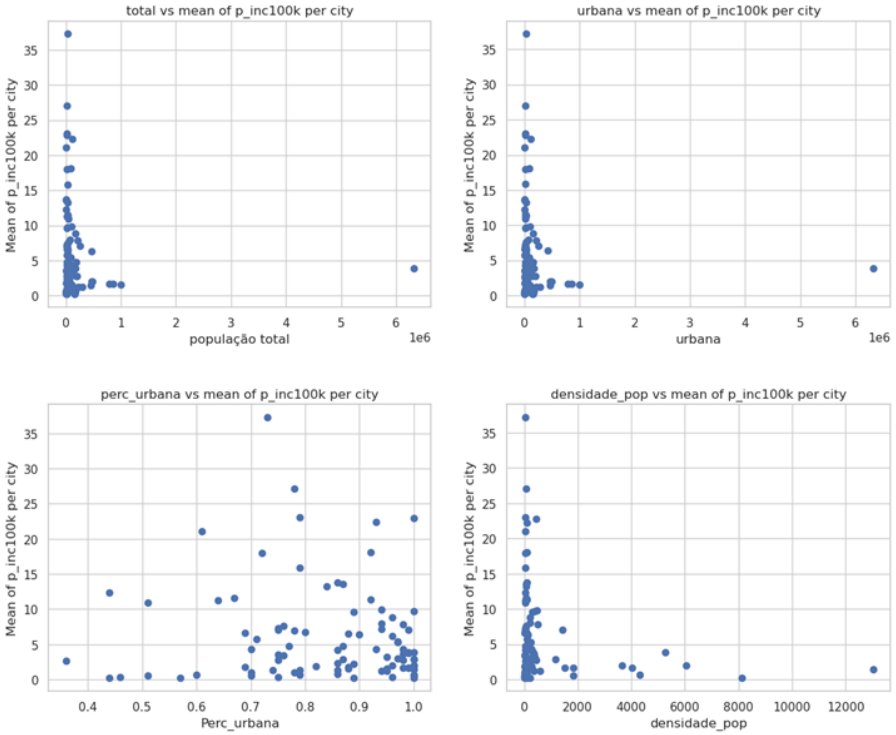

Figura 20 - Dispersão entre variáveis demográficas e incidência média municipal

	População	População urbana	% população urbana	Densidade populacional
População	1	1	0,21	0,38
População urbana	1	1	0,22	0,38
% população urbana	0,21	0,22	1	0,34
Densidade populacional	0,38	0,3	0,34	1

Figura 21 - Matriz de correlação das variáveis demográficas selecionadas

A partir dessas análises, nota-se que, ao contrário do que se imaginava, o percentual da população urbana é a variável menos correlacionada com a incidência média de dengue municipal. Além disso, notamos que a população urbana possui correlação de postos máxima com a população total e, portanto, não faz sentido manter ambas as variáveis em um modelo de regressão.

Sendo assim, mantivemos apenas as variáveis que se referem à população total e à densidade populacional.

ii. Engenharia de variáveis

Nesta seção, é descrita a criação de algumas novas variáveis que ajudariam a explicar o comportamento da curva de casos de dengue nos municípios do RJ a partir de transformações de variáveis já existentes.

Partindo de conhecimento de modelos autorregressivos em séries temporais, foi criada uma variável que representa o número de casos duas semanas antes. Ou seja, com uma defasagem (‘lag’) de 14 dias. Essa variável foi denominada de casos_lag14. Pensando em um modelo capaz de prever a quantidade de casos de dengue ou a incidência deles, era de se imaginar que o número de casos duas semanas antes ajudaria a predizer o número de casos atual. Naturalmente, poderia ser utilizado uma defasagem de uma semana no lugar de duas, mas isso prejudicaria a aplicabilidade do modelo devido à um menor período de antecipação para se planejar.

A segunda nova variável foi o mês, obtido a partir da data de início da semana epidemiológica. Apesar de simples e um tanto quanto trivial, essa variável foi importante para denotar a sazonalidade anual que ocorria na curva de casos de dengue no estado do RJ. Como podemos observar na Figura 22 e no gráfico da figura 10, alguns meses do ano apresentam um número de casos muito maior que os demais. Neste sentido, destacam-se os meses de março à maio, seguidos fevereiro e junho.

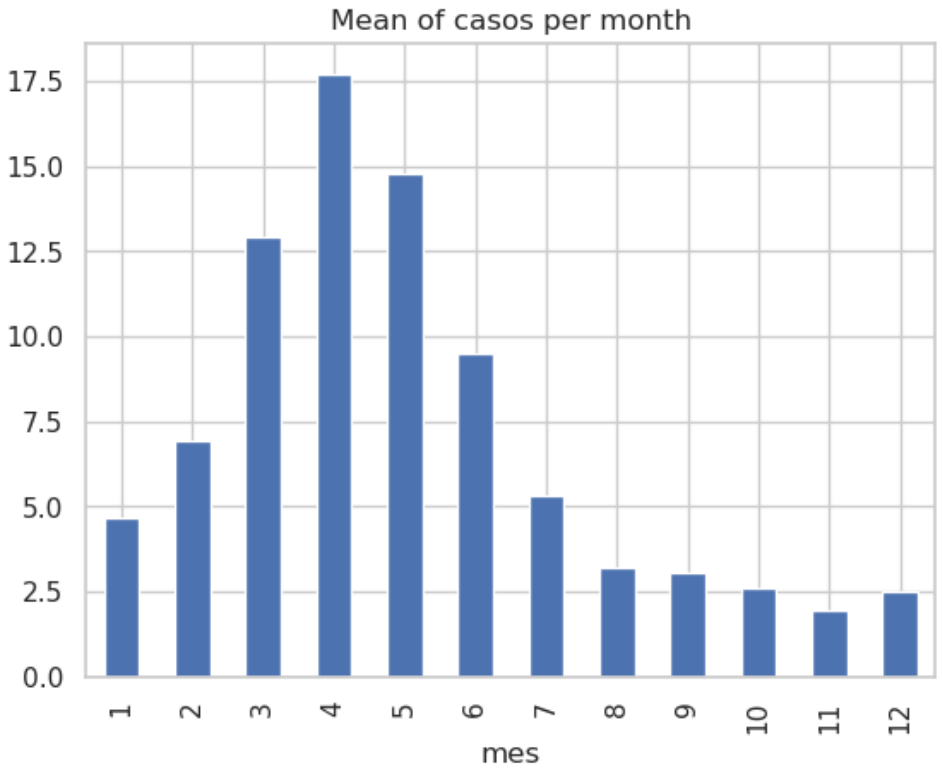

Figura 22 - Média de casos por mês

Nesta mesma linha, analisou-se o a média de casos por estação do ano. Para simplificações, consideramos a seguinte relação:

- Verão: janeiro, fevereiro e março
- Outono: abril, maio e junho
- Inverno: julho, agosto e setembro
- Primavera: outubro, novembro, dezembro

Como é possível ver nas Figura 23 e 24, e em decorrência do comportamento mensal já visto no gráfico anterior, as estações de outono e verão se destacam em número de casos, apresentando uma média de casos amplamente superior que a primavera e o inverno.

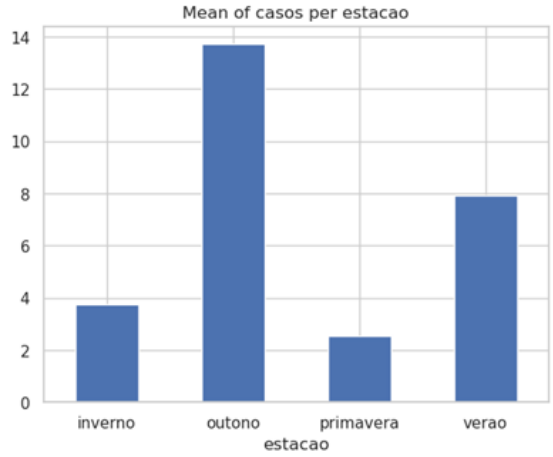

Figura 23 - Média de casos por estação

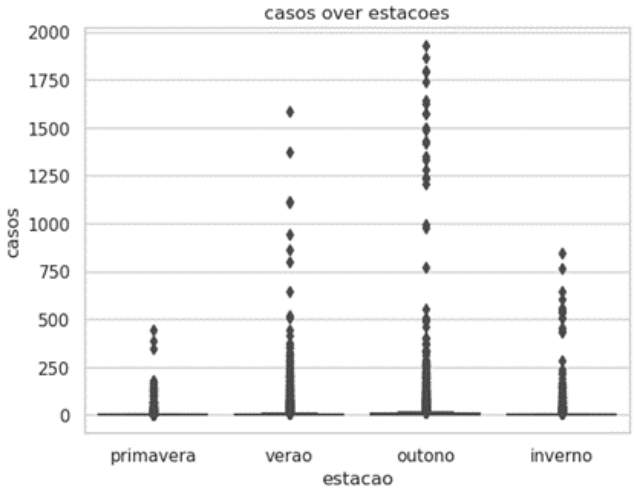

Figura 24 - Boxplot de casos por estação

Tendo isso em vista e visando diminuir a quantidade de categorias da variável, agrupou-se as estações em apenas duas: estação de alta (outono e verão) e estação de baixa (primavera e inverno). A derivação obtida pode ser vista na Figura 25.

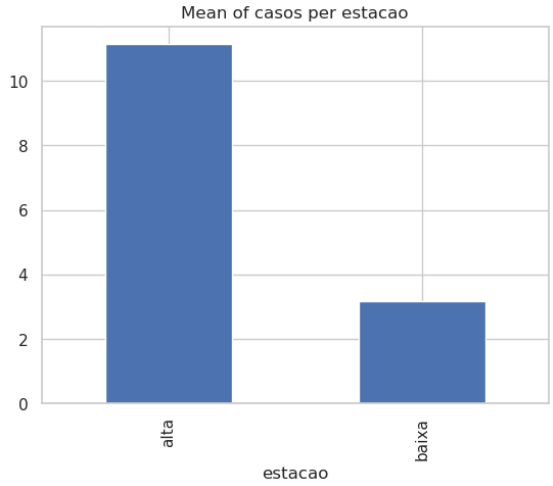

Figura 25 - Média de casos por estação agrupada

Para checar se havia diferença de casos entre os dois grupos no sentido estatístico, foi feito um teste de Shapiro que apontou para a não normalidade da distribuição da variável. Portanto, foi realizado um teste não paramétrico de Mann-Whitney, que confirmou que a diferença no número de casos entre as duas estações é estatisticamente relevante.

Seguindo pela análise das variáveis na linha do tempo, um maior entendimento do processo que se dá desde a reprodução do mosquito até a notificação de um caso de dengue foi importante para reformular a modo como se estava trabalhando com os fatores climáticos.

O mosquito Aedes Aegypti, como mencionado, é o principal vetor da doença e sua reprodução é fortemente impactada pelas características climáticas. Os ovos do mosquito são depositados em poços de água parada, algo que se torna mais disponível após as chuvas. Na sequência, o desenvolvimento da larva até a fase adulta do mosquito, na qual sua atividade aumenta, depende de condições de temperatura e umidade adequadas, mas dura em média entre 7 e 10 dias (FIOCRUZ, 2019). Somando o tempo desde o acasalamento e desenvolvimento da larva até a fase adulta do mosquito, totaliza um período entre 14 e 21 dias. Por isso, fazia sentido trabalhar com as variáveis de temperatura, umidade e precipitação defasadas desse período.

Levando isso em consideração, foram criadas duas variáveis denominadas tempmed_lag21 e precipit_sum_lag21, que representavam a temperatura média e a precipitação total 3 semanas antes, sinalizando para condições propícias de desenvolvimento do mosquito.

Um mosquito em atividade se torna um transmissor da doença cerca de 10 a 12 dias após picar uma pessoa acometida por dengue (VALLE, s.d.), sendo que, após picada, o vírus fica encubado de 2 a 10 dias até que os sintomas comecem a aparecer (Secretaria de Estado de Saúde do Mato Grosso, s.d.).

No total, são de 12 a 21 dias entre um mosquito picar alguém infectado, picar um novo indivíduo e essa pessoa notificar. De maneira simplificada, isso significa que os casos de hoje impactam no número de casos de 2 a 3 semanas depois e, portanto, esse comportamento já estava sendo expresso pela variável casos_lag14.

Vale considerar também que a atividade do mosquito é fortemente dependente da temperatura e precipitação, sendo máxima em temperaturas entre 25º e 30º Celsius.

Levando essa perspectiva qualitativa em consideração e também analisando os comportamentos quantitativamente, foram propostas algumas novas variáveis:

- Tempmed_lag21: temperatura média da semana epidemiológica iniciada 21 dias antes
- Precipitation_sum_ms_lag14: precipitação total da semana epidemiológica iniciada 14 dias antes
- Precipitation_sum_ms_lag21: precipitação da semana epidemiológica iniciada 21 dias antes
- Umidade_lag21: umidade média da semana epidemiológica iniciada 21 dias antes

Abaixo (Figura 26) se encontram os gráficos de dispersão dessas variáveis versus a incidência. Dentre eles, o que mais chama atenção é o de tempmed_lag21.

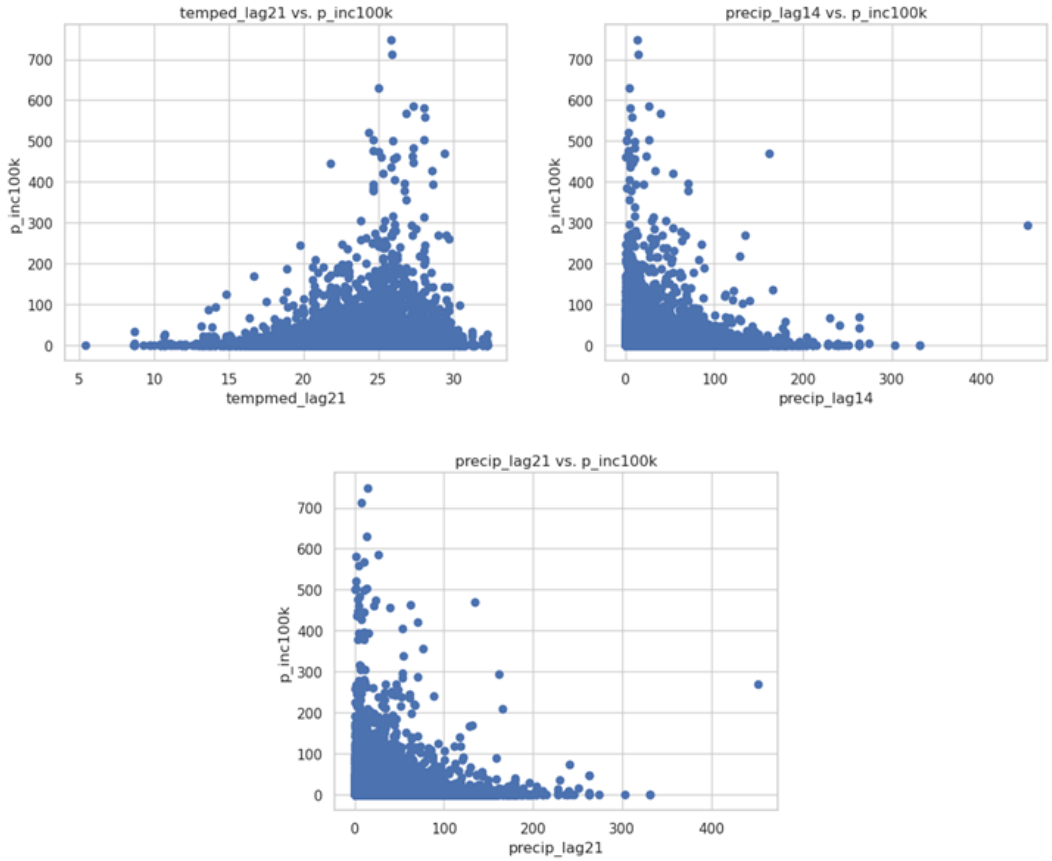

Figura 26 - Dispersão entre variáveis climáticas defasadas e incidência

Ademais, o gráfico da Figura 27 sinalizou para uma possível antecipação da sazonalidade mensal de casos pela temperatura média de meses anteriores. Nele, podemos ver a média mensal de casos (barras) e comparação com a temperatura média desses meses (linha). Em contrapartida, se adiantamos a temperatura em 2 meses, as curvas parecem se acompanhar surpreendentemente bem (veja Figura 28).

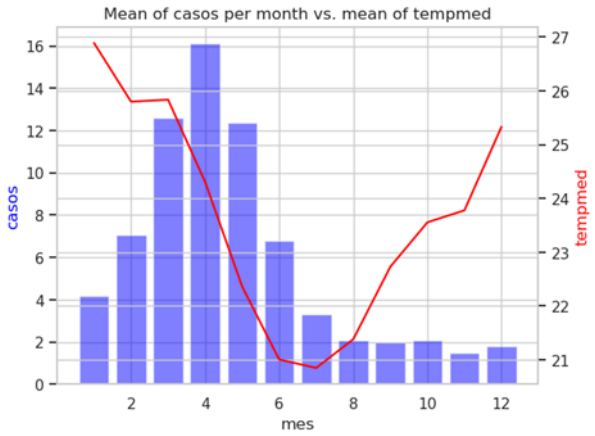

Figura 27 - Média de casos mensal versus média de temperatura

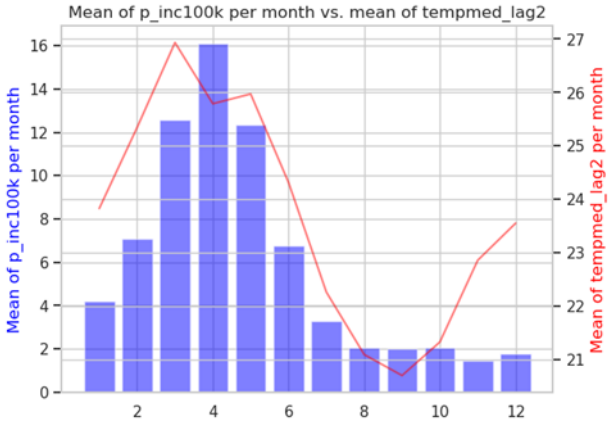

Figura 28 - Média de casos mensal versus média de temperatura defasada de 2 meses

Tendo isso em vista, foi criada uma nova variável que representava a temperatura média acumulada mensal de 2 meses antes, denominada tempmed_lag2. Para confirmar que essa variável era relevante, foram calculadas as correlações de Pearson e Spearman em relação à incidência de dengue, encontrando um resultado de 0,120 e 0,130 respectivamente. Por mais que não sejam valores muito expressivos, ao menos acenam para uma contribuição que essa variável possa desempenhar no modelo.

A mesma análise foi repetida para a precipitação e as conclusões foram as mesmas, conforme podem ser vistas nos gráficos abaixo. A Figura 29 representam as curvas reais e a Figura 30 apresenta a curva de precipitação total acumulada mensal adiantada de 2 meses.

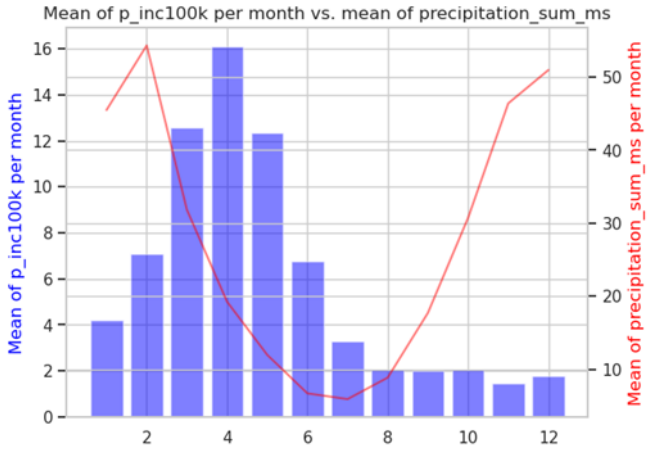

Figura 29 - Média de casos mensal versus média de precipitação

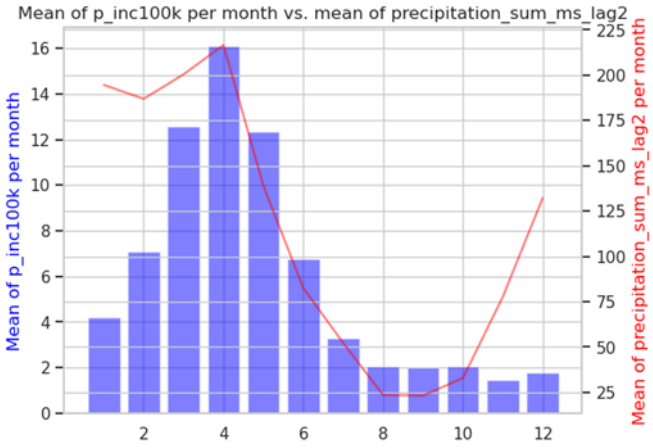

Figura 30 - Média de casos mensal versus média de precipitação defasada de 2 meses

Analogamente, foi criada a variável precipitation_sum_ms_lag2, que representava a precipitação total acumulada mensal de 2 meses antes. As correlações de Pearson e Spearman também foram calculadas, encontrando respectivamente os valores de 0,099 e 0,080. Por menos expressivas que fossem essas correlações, optou-se por manter tal variável em função da análise gráfica anterior.

A título de conhecimento, tentou-se repetir a mesma lógica para a umidade, mas não foi possível chegar às mesmas conclusões (veja Figura 31).

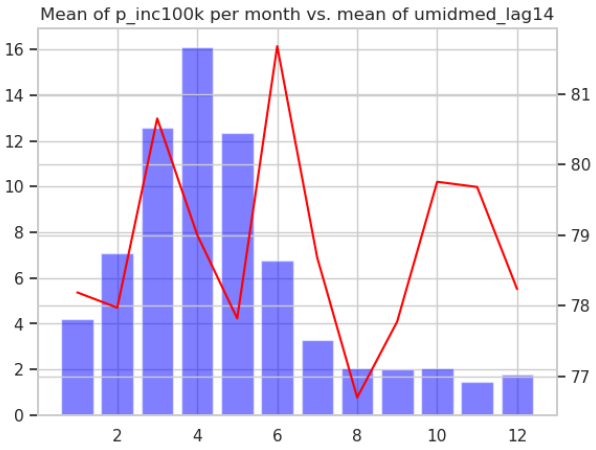

Figura 31 - Média de casos mensal versus média de umidade

iii. Base final

Tendo feito as etapas de tratamento dos dados e criação das novas variáveis, era possível finalmente consolidar a base final. Em razão das variáveis defasadas, tivemos uma perda de dados de algumas semanas do começo do período analisado. Entretanto, a perda de informação foi desprezível se comparada a todo o período analisado.

Ao fim, a base era composta por 24.281 linhas e 23 colunas. Dessas 23 colunas, 3 eram variáveis de apoio (se referiam ao código e nome do município e data de início da semana epidemiológica), 18 eram variáveis explicativas do modelo e 2 eram possíveis variáveis respostas – casos de dengue e incidência de casos de dengue.

Dentre as variáveis explicativas, 3 delas eram categóricas e se referiam à receptividade do clima, mês e estação. Para fins de treinamento do modelo, é importante que as categorias tenham volume suficiente de dados, o que pode ser constatado através da Figura 32.

		RECEPTIVO	
ESTAÇÃO	MÊS	0	1
alta	1	1143	748
	2	1217	967
	3	1276	1090
	4	1697	602
	5	2225	145
	6	1868	20
baixa	7	1999	0
	8	1977	5
	9	1834	106
	10	1746	198
	11	1550	270
	12	1186	412

Figura 32 - Volume de dados por combinação de categorias

Quanto às variáveis numéricas, foram analisadas as correlações entre as variáveis explicativas para checar uma possível multicolinearidade. Como é possível averiguar no mapa de calor da Figura 33, que representa a matriz de correlação, a maior correlação entre duas variáveis explicativas distintas é de 0,71, demonstrando não haver nenhuma situação crítica de multicolinearidade.

É importante frisar que as duas últimas colunas e linhas são as possíveis variáveis respostas e não há problema de haver alta correlação entre alguma delas e as variáveis explicativas.

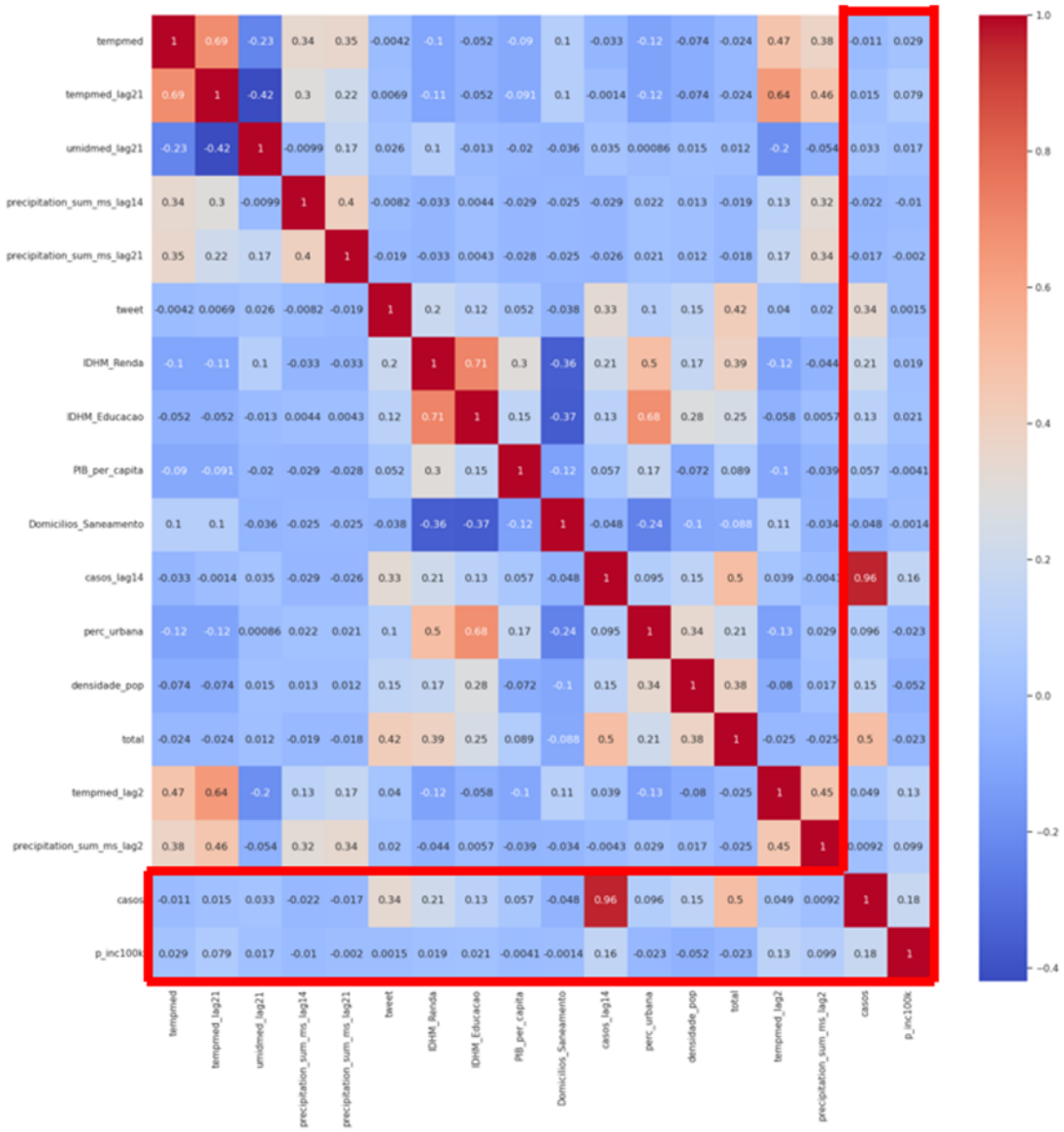

Figura 33 - Matriz de correlação das variáveis numéricas do modelo

c. Construção do modelo

i. Definição da variável resposta

Tendo em mãos a base final tratada e as variáveis explicativas do modelo consolidadas, restava ainda uma definição essencial para o treinamento preditivo: a variável resposta.

Conforme mencionado anteriormente, estavam sendo consideradas até então a variável número de casos e incidência de casos por 100mil habitantes.

Para isso, é importante levar em consideração a discrepância existente entre a população dos municípios. Por exemplo, o Rio de Janeiro, município mais populoso do estado, tem uma população de 6.320.446 habitantes segundo o censo de 2010, enquanto o município de Macuco, menos populoso da região, tem apenas 5.269 habitantes. Da mesma forma, a média de casos no Rio de Janeiro foi de 237,34 casos por semana, enquanto a de Macuco foi de míseros 0,38.

Em contrapartida, se olhamos para incidência ao invés do número absoluto de casos, Macuco apresenta uma incidência média quase duas vezes maior que a do Rio de Janeiro – Macuco teve uma média de 7,21 casos/100mil habitantes, enquanto o Rio de Janeiro teve 3,76. Essa análise pode ser vista na Figura 34.

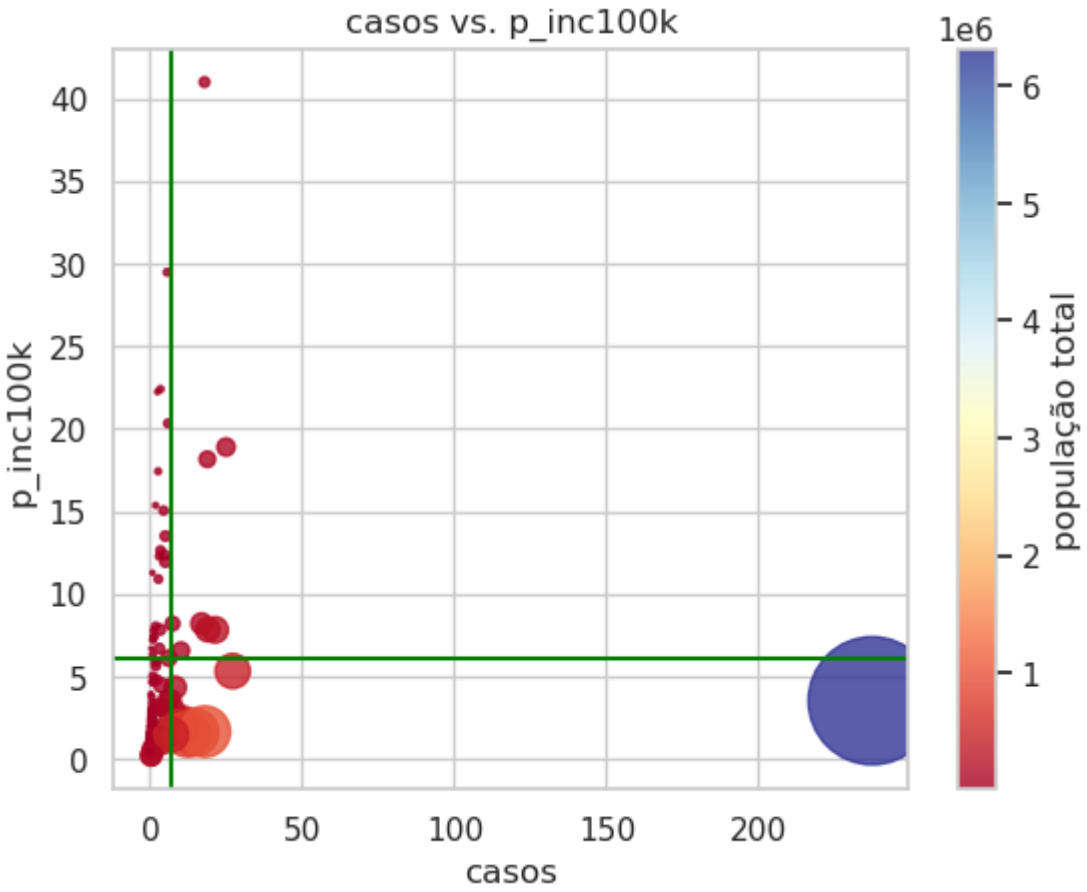

Figura 34 - Dispersão entre média de casos e incidência municipal

Em linhas práticas, isso significa que, ao adotar a incidência como variável resposta, o modelo daria menos peso aos casos ocorridos em municípios mais populosos. Ou seja, os casos seriam ponderados pelo inverso da população.

Por mais que a utilização da incidência como variável de análise seja amplamente adotado em estudos da área de saúde (sobretudo em epidemiologia), no caso do presente trabalho, isso poderia levar a erros maiores na previsão de casos de municípios populosos e críticos para o sistema de saúde. Sendo assim, e pensando na usabilidade deste modelo para apoiar políticas públicas, optou-se por utilizar o número absoluto de casos como variável dependente, expurgando da base a variável p_inc100k.

ii. Pré-processamento da base

Na sequência, as bases foram divididas em dados de treinamentos e dados de teste aleatoriamente em uma proporção de 80/20.

A fim de deixar as variáveis em formatos adequados para serem utilizadas por qualquer algoritmo de Machine Learning, foram realizados alguns pré-processamentos.

O primeiro deles foi a conversão das variáveis categóricas em variáveis dummy (ou variáveis indicadoras). Esse processo é particularmente importante quando se trabalha com modelos de aprendizado de máquina e análise de regressão, que geralmente requerem entradas numéricas. Ao fazer essa conversão, a variável categórica é transformada em N-1 variáveis binárias, onde N é o número de categorias da variável original. Cada variável binária representa a presença ou ausência de determinada categoria, mas uma das categorias é desconsiderada pois pode ser determinada através de relações lineares entre as demais, evitando assim multicolinearidade.

As variáveis numéricas foram normalizadas em seguida. Para isso, utilizou-se a padronização pelo z-score de cada variável – considerando os parâmetros estatísticos apenas dos dados de treinamento. O z-score pode ser calculado subtraindo-se a média e dividindo pelo desvio-padrão, conforme fórmula abaixo. Fazendo isso, a variável terá sua média em zero e desvio padrão igual a 1, o que é importante para algoritmos que assumem que os dados são normalmente distribuídos, como a regressão linear. Além disso, para esse algoritmo, também facilita a interpretação dos coeficientes das variáveis.

É importante mencionar também que o mesmo pré-processamento aplicado na base de treinamento foi reproduzido na base de teste. Isso é, para os casos de transformação em dummies, eliminou-se a coluna binária referente à mesma categoria determinada na transformação da base de treinamento. Já no caso das variáveis numéricas, a normalização foi feita considerando a média e desvio padrão dos dados de treinamento.

Além disso, deve-se clarificar que a variável resposta – que nesse caso era numérica – permanece inalterada.

iii. Otimização de hiperparâmetros

Concluído o pré-processamento dos dados, foram definidos os algoritmos de regressão os quais seriam treinados para predizer o número de casos por semana e município e posteriormente comparados. Visando previsões de qualidade e menor custo computacional, foram escolhidos os seguintes algoritmos: Regressão Linear, Árvore de Decisão, Extra Gradient Boosting e Random Forest.

Entretanto, antes de treinar os modelos e comparar os resultados na base de validação, era necessário fazer o ajuste – ou melhor dizendo, otimização – de hiperparâmetros.

A otimização de hiperparâmetros é uma etapa crucial no desenvolvimento de modelos de aprendizado de máquina. Hiperparâmetros são as configurações ajustáveis de um algoritmo que não são aprendidas a partir dos dados, mas sim definidas por quem o está regendo, e têm um impacto significativo no desempenho do modelo treinado.

No entanto, encontrar o conjunto ideal de hiperparâmetros não é uma tarefa trivial. É aqui que entram técnicas como a validação cruzada (cross-validation), K-Fold e Grid Search.

A técnica da validação cruzada envolve a divisão do conjunto de dados em várias partes, ou "folds". No caso deste trabalho, foram utilizadas divisões em 5 partes. O modelo é treinado em 4 destas partes e validado na parte restante. Isso é repetido várias vezes, de modo que cada parte do conjunto de dados é usada tanto para treinamento quanto para validação. Esta abordagem ajuda a garantir que o modelo seja robusto e generalizável para novos dados. Ou seja, evitar que as escolha dos hiperparâmetros seja boa apenas para o conjunto de dados específico de treinamento.

Entretanto, considerando a quantidade de combinações diferentes de hiperparâmetros e a quantidade de folds escolhidos na validação cruzada, o número de iterações necessárias para completar a avaliação dos hiperparâmetros pode ser muito alto. Multiplicando esse número pela complexidade computacional dos algoritmos que serão utilizados para treinar os modelos, resulta-se um tempo de processamento muito alto.

Neste contexto, algumas técnicas, como Randomized Grid Search, são aplicadas para reduzir esse custo computacional. Esta técnica envolve a criação de uma "grade" de hiperparâmetros e a seleção aleatória de combinações de hiperparâmetros para testar. Diferentemente do Grid Search tradicional, que testa todas as possíveis combinações de hiperparâmetros, o Randomized Grid Search seleciona aleatoriamente um subconjunto destas combinações, economizando tempo e recursos computacionais.

No caso dos algoritmos testados neste trabalho, a Figura 35 denota o conjunto de hiperparâmetros e seus valores testados no processo de validação cruzada.

Algorítmos e hiperparâmetros avaliados

Linear Regression	Decision Tree	XGBoost	Random Forest
fit_intercept	**max_depth**	**n_estimators**	**bootstrap**
[True, False]	[None, 5, 10]	[100, 150, 200]	[False, True]
	min_samples_leaf	**max_depth**	**max_depth**
	[1, 2, 4]	[5, 10, 30, 50]	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]
	max_features	**learning_rate**	**max_features**
	['auto', 'sqrt', 'log2']	[0.1, 0.01, 0.001]	['auto', 'sqrt', 'log2', 'None']
	random_state	**gamma**	**min_samples_leaf**
	[10]	[0, 0.1, 0.2]	[1, 2, 4]
		random_state	**min_samples_split**
		[10]	[2, 12, 22, 32, 42, 52]
			n_estimators
			[50, 100, 150, 200, 250]

Figura 35 - Hiperparâmetros avaliados para os modelos

Avaliar a performance de modelos de regressão é crucial para entender o quão bem um modelo está prevendo os valores reais. Essa métrica é utilizada no processo de GridSearch para definir qual foi o conjunto de hiperparâmetros que apresentou melhor resultado e a escolha cuidadosa da métrica adequada é essencial para garantir que os hiperparâmetros apontados pela validação cruzada resultem em um modelo adequado para os propósitos do trabalho.

A métrica escolhida para avaliar a qualidade dos modelos na validação cruzada foi a MSE (Mean Squared Error), que é calculada pela média dos quadrados das diferenças entre o valor real e o valor preditos pelo modelo. Outra métrica comumente usada é a MAE (Mean Abslute Error), que significa erro médio absoluto.

Comparando ambas, vale lembrar que a primeira penaliza mais erros maiores, o que significa que o MSE é mais sensível à outliers do que o MAE. Em um cenário onde é crucial prever assertivamente aumentos abruptos de caso e não há margem para grandes erros de planejamento, é necessário que esse comportamento seja refletido no modelo. Por tal razão, foi utilizado o MSE como principal métrica de avaliação de desempenho do modelo.

É importante observar também que métricas de erros relativos, como o MRAE (Mean Relative Absolute Error), seriam um problema em uma variável dependente inflada de zeros.

Ao fim da validação cruzada, foram apontados os melhores hiperparâmetros dos modelos e os respectivos MSEs mensurados, que podem ser vistos na tabela abaixo.

Algoritmo	MSE	Parâmetros
Regressão Linear	246,05	{'fit_intercept': False}
Árvore de Decisão	1140,34	{'random_state': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 5}
Extra Gradient Boosting	2302,54	{'random_state': 10, 'n_estimators': 100, 'max_depth': 30, 'learning_rate': 0.001, 'gamma': 0.1}
Random Forest	764,40	{'random_state': 10, 'n_estimators': 250, 'min_samples_split': 52, 'min_samples_leaf': 4, 'max_features': 'log2', 'max_depth': 30, 'bootstrap': True}

Tabela 4- Melhores hiperparâmetros por modelo e respectivos scores

Como é possível analisar na Tabela 4, a Regressão Linear foi o modelo que melhor performou, com um MSE de 246,05. O segundo melhor modelo, o Random Forest, teve um MSE de 764,40, mais que 3 vezes superior ao da regressão.

iv. Escolha do modelo

Tendo em vista os resultados ruins dos algoritmos de Árvore de decisão de Extra Gradient Boosting na validação cruzada, optou-se por delimitar os modelos que continuariam a ser avaliados. Assim, apenas a Regressão Linear e o Random Forest seguiram sendo avaliados no intuito de definir o algoritmo que seria adotado para o modelo final.

Para que se tivesse mais métricas para comparação dos modelos, eles foram treinados na base de treinamento completo utilizando os hiperparâmetros ótimos de acordo com a validação cruzada. Feito isso, os modelos foram aplicados em cima do conjunto de dados de validação e o número de casos predito foi comparado ao número de casos real já conhecido. A partir desse resultado, foi possível calcular três das principais métricas de avaliação de modelos de regressão: MAE, MSE e R^2.

Esta última métrica representa a proporção da variância na variável dependente que é previsível a partir das variáveis independentes. Em termos mais simples, ele indica a porcentagem da variação da variável de resposta que é explicada pelo modelo. Para o R^2, quanto maior melhor, ao contrário das outras duas métricas.

As avaliações dos dois modelos estão representadas na Tabela 5 abaixo.

Algoritmo	MAE	MSE	R^2
Regressão Linear	4,15	311,69	0,89
Random Forest	4,60	744,99	0,74

Tabela 5 - Métricas de desempenho dos dois melhores modelos

Como podemos ver na tabela acima, Regressão Linear performa melhor segundo todas as métricas, especialmente MSE e R^2.

v. Correção de previsões negativos

Apesar de, novamente, a regressão linear se mostrar o melhor algoritmo para a previsão de casos de dengue no RJ, havia ainda um detalhe que não estava sendo considerado. Nenhum dos modelos estava levando em conta a restrição de que o número de casos previsto não poderia ser inferior a zero. Ou seja, em muitos casos, o modelo previa um número negativo de casos, o que é um absurdo do ponto de vista prático.

Para corrigir esse problema e melhorar ainda mais o desempenho do modelo, aplicou-se um mapeamento na saída do modelo para corrigir valores negativos para zero. Desta forma, essa tratativa da previsão atuava como um segundo sistema acoplado em série ao sistema representado pelo modelo preditivo inicial.

O resultado final de cada um dos algoritmos adicionado o mapeamento de valores negativos pode ser visto na Tabela 6.

Algoritmo	MAE	MSE	R^2
Regressão Linear	3,47	309,49	0,89
Random Forest	4,18	584,03	0,80

Tabela 6 - Desempenho dos modelos após limite inferior da saída em zero

O que podemos notar em relação ao resultado do sistema anterior é que houve uma redução considerável do MAE para ambos os algoritmos. No entanto, não houve uma queda significativa do MSE da regressão linear. Isso pode ser explicado pelo fato de os mapeamentos de valores negativos feitos terem melhorado previsões cujo erro não era grande, mas que impactaram pouco uma métrica que penaliza erros de amplitude maior.

Entretanto, a queda significativa do MSE do Random Forest leva a entender que as previsões de valor negativo desse modelo eram de grande amplitude e, portanto, também eram os seus erros, que foram minimizados pelo mapeamento.

Não obstante esse aperfeiçoamento do Random Forest, os resultados ainda mostram que o melhor modelo foi o em que se utilizou a regressão linear como algoritmo preditor. Esta foi a escolha final para o trabalho.

4. Resultados e discussão

O modelo discutido na seção anterior apresentou uma assertividade bastante satisfatória, mas é interessante analisar outras estatísticas a respeito dele para se ter uma visão mais completa. Essas estatísticas podem ser visualizadas na Figura 36 e remetem aos resultados do modelo na base de treinamento.

Métrica	Valor
Variável Dependente	casos
R-quadrado	**0,917**
R-quadrado Ajustado	0,916
Método	Least Squares
Estatística F	**7895**
Prob (F-estatística)	**0**
Log-Likelihood	-80218
Nº Observações	19424
DF Residuais	19397
DF Modelo	27
AIC	**1,61E+08**
BIC	**1,61E+08**
Tipo de Covariância	nonrobust

Figura 36 - Avaliação dos resultados da Regressão Linear

Como se pode ver, o modelo tem um R-quadrado e um R-quadrado ajustado muito altos, com valor de 0.917 e 0.916 respectivamente. Isso indica que aproximadamente 91.6% da variabilidade da variável dependente 'casos' pode ser explicada pelas variáveis independentes no modelo na base de treinamento. Aqui é importante ressaltar que esse valor diverge do mostrado na tabela Y pois, enquanto o primeiro foi mensurado na base de validação, este novo valor reflete o desempenho na base de treinamento.

Ademais, o alto valor da estatística F (7815) com um p-valor de praticamente zero sugerem que o modelo é estatisticamente significativo. Isto é, as variáveis independentes, coletivamente, têm um efeito significativo sobre a variável dependente.

O número de observações é 19424, o que é uma amostra robusta, proporcionando confiança nos resultados do modelo. Os graus de liberdade residuais são 19397, que é calculado subtraindo-se a número de variáveis independentes no modelo (27) do número de observações.

AIC e BIC são medidas de qualidade do modelo que penalizam a complexidade do modelo, e seus valores relativamente altos sugerem que o modelo pode ser complexo, o que é esperado dado o alto número de preditores. Isso indica a possibilidade de sobre ajuste e baixa significância de algumas variáveis independentes.

Apesar de todas as descrições estatísticas do modelo, a visualização gráfica colabora para a compreensão dos resultados. Os gráficos das Figuras 37 e 38, por exemplo, exibem a quantidade de casos preditos (azul) em comparação aos realizados (laranja), considerando apenas as amostras de validação. Como se pode ver, a performance é boa se levando em conta todos os municípios e a distribuição desses casos no decorrer do ano.

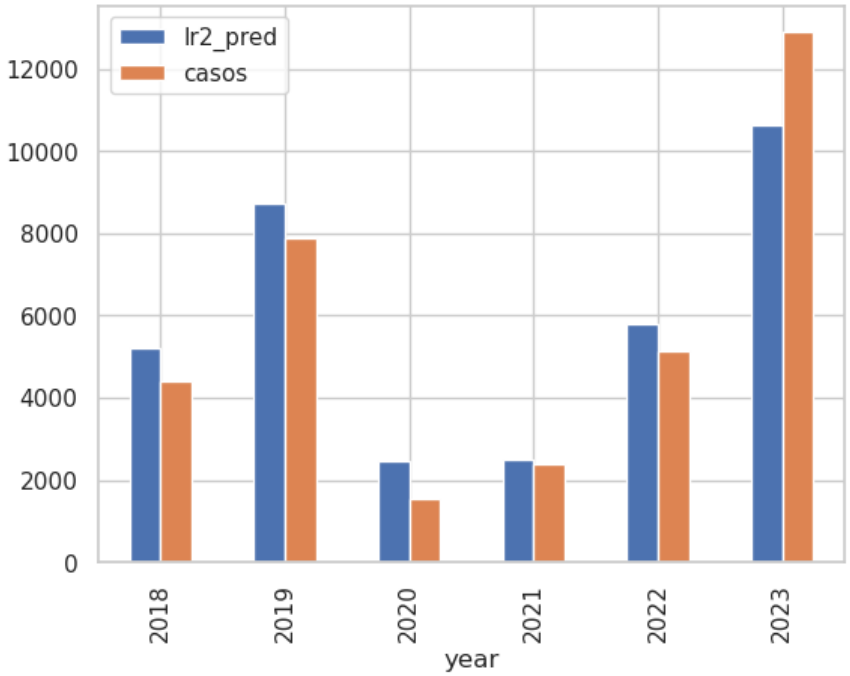

Figura 37 - Comparação entre previsto e realizado por ano

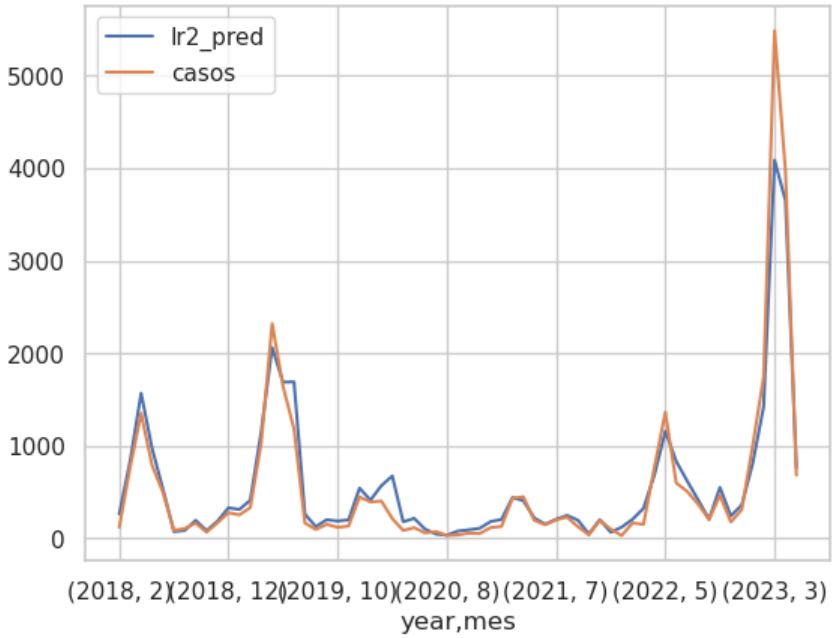

Figura 38 - Séries temporais de casos previstos e realizados

Todavia, é possível notar um erro absoluto maior de previsão no ano de 2023, que exige uma análise mais detalhada para entender onde esse erro está localizado: em qual município e em qual semana.

Dando um enfoque maior no ano de 2023 (veja Figura 39), podemos ver que esse erro é predominante nas semanas que vão do dia 12 até 26 de março, quando ocorre um abrupto crescimento do número de casos. Apesar do erro, o modelo mesmo assim foi capaz de prever essa alta súbita de casos no mês de março, o que, para fins de planejamento das organizações de saúde, já seria alarmante o suficiente.

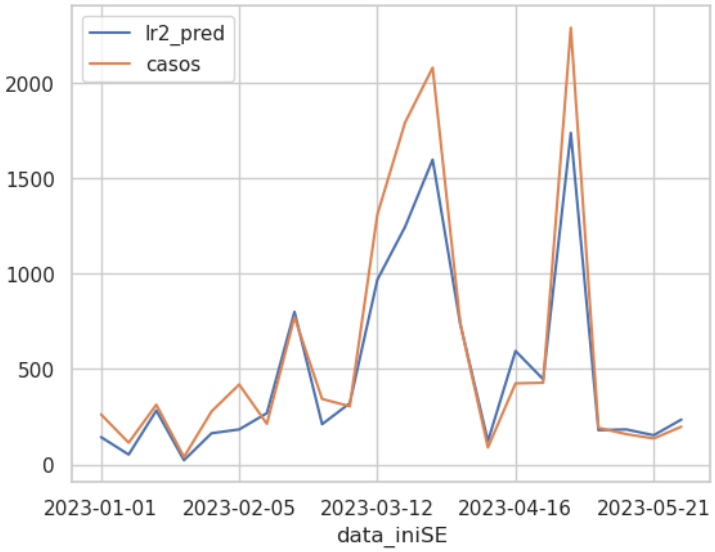

Figura 39 - Séries temporais de casos previstos e realizados em 2023

Quebrando por município nesse mesmo ano, podemos notar a predominância desse erro na cidade do Rio de Janeiro, conforme a Figura 40.

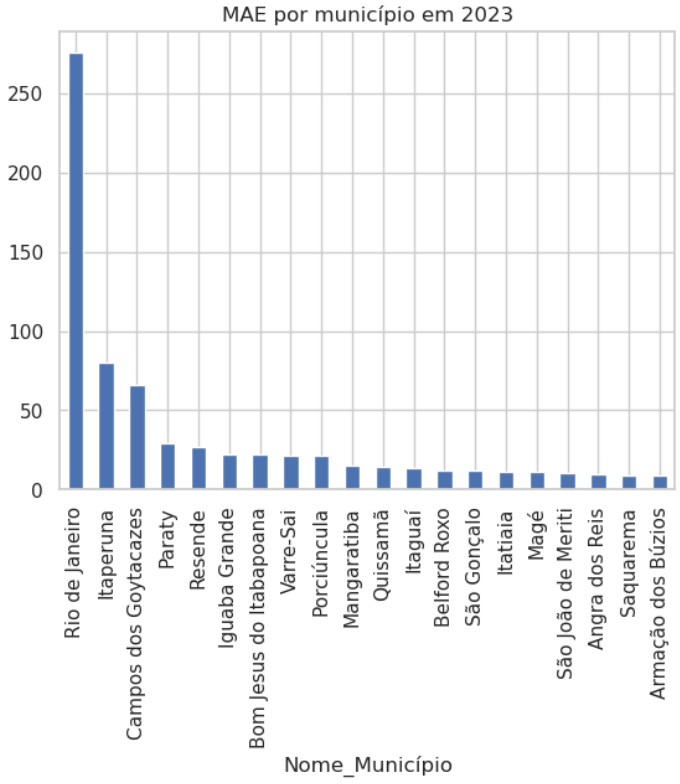

Figura 40 - Erro médio absoluto por município

Quando analisada do ponto de vista da localidade em todo o horizonte de tempo, os erros são bem distribuídos. Chama a atenção positivamente a alta assertividade das previsões nos municípios mais afetados, que podem ser vistos na Figura 41.

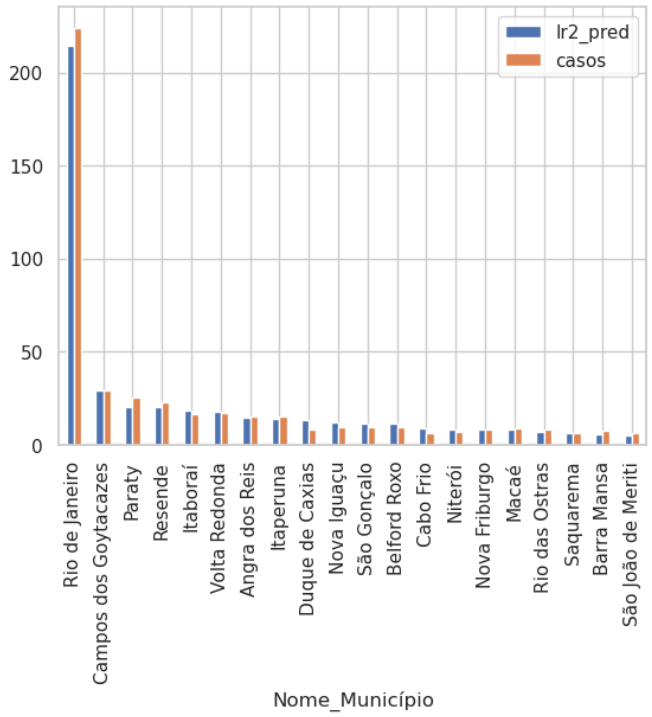

Figura 41 - Comparação entre previsto e realizado para os 20 municípios mais afetados

Em compensação, ao se analisar os erros nos municípios com menor número de casos (Figura 42), podemos notar que o modelo de forma geral sobre-estima o número de casos, o que não necessariamente é crítico se observado que o número de casos estimado não é alto, sendo menos que 1 caso em média para a maior parte desses municípios. Assim, em termos de tomada de decisão, esse erro não deve ser impactante.

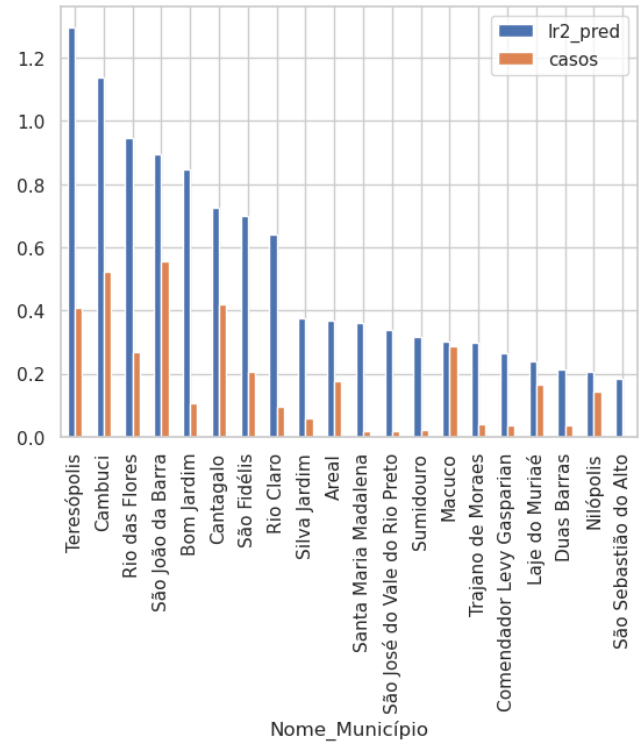

Figura 42 - Comparação entre previsto e realizado para os 20 municípios menos afetados

Para reforçar a conclusão de que o maior volume de erros ocorre em municípios com menor número de casos, as previsões para cada município foram plotadas em um gráfico de dispersão em função do número de casos total, na Figura 43. Ou seja, no eixo x, temos o número de casos total da base de teste por município e no eixo y a soma das previsões por município.

O tamanho dos pontos representa o RMAE (Relative Mean Absolute Error), que é calculado dividindo-se a soma dos erros absolutos pela soma do número de casos do município. Essa abordagem foi uma alternativa para relativizar os erros a nível municipal evitando possíveis divisões por zero em registros onde o número de casos é nulo. Já as cores denotam o erro absoluto total.

A bissetriz traçada em verde no meio do gráfico representa a situação de otimalidade na qual o número de casos previsto é igual ao número de casos reais.

O segundo gráfico (Figura 44) é apenas uma amplificação do primeiro desconsiderando o município do Rio de Janeiro.

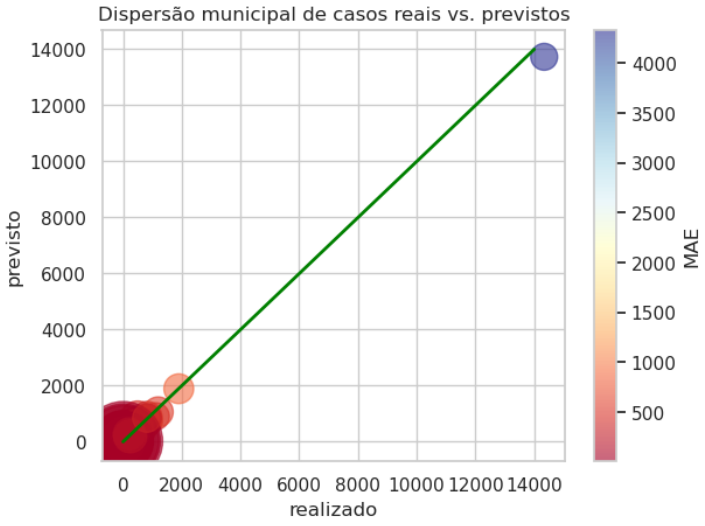

Figura 43 - Dispersão realizado vs. previsto por município

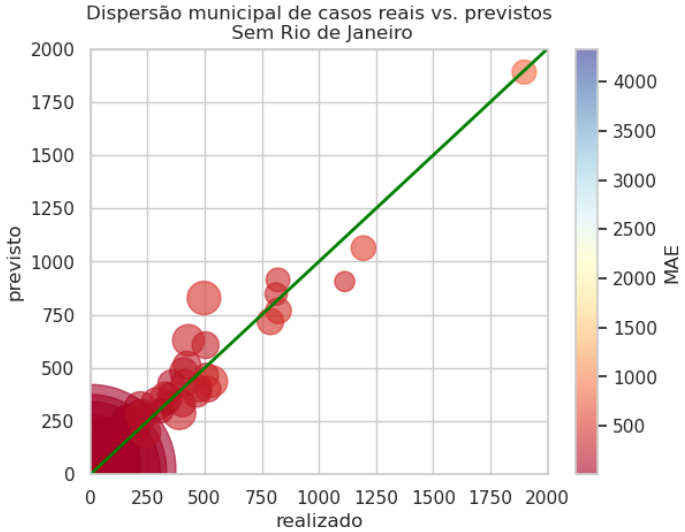

Figura 44 - Dispersão realizado vs. previsto por município sem Rio de Janeiro

Como é possível notar, os municípios com maior erro absoluto se encontram no canto inferior esquerdo, significando que tiveram menor número de casos realizados. Logo, pequenos erros de previsão acabam sendo significativos se comparados com o número médio de casos no local.

Entretanto, olhando explicitamente para o erro absoluto total, o pior índice se encontra no canto superior direito. Mais especificamente, na cidade do Rio de Janeiro. Como já mencionado anteriormente, contudo, esses erros são desprezíveis frente ao volume total de casos que ocorrem no município mais populoso do estado, que pode ser visto pela proximidade com a bissetriz.

Para tornar a avaliação mais completa e a visualização mais detalhada, foram feitas previsões em cima da base de treinamento também. Juntando as previsões da base de teste e treinamento e comparando-as com o número de casos reais na linha do tempo, chegamos ao gráfico da Figura 45.

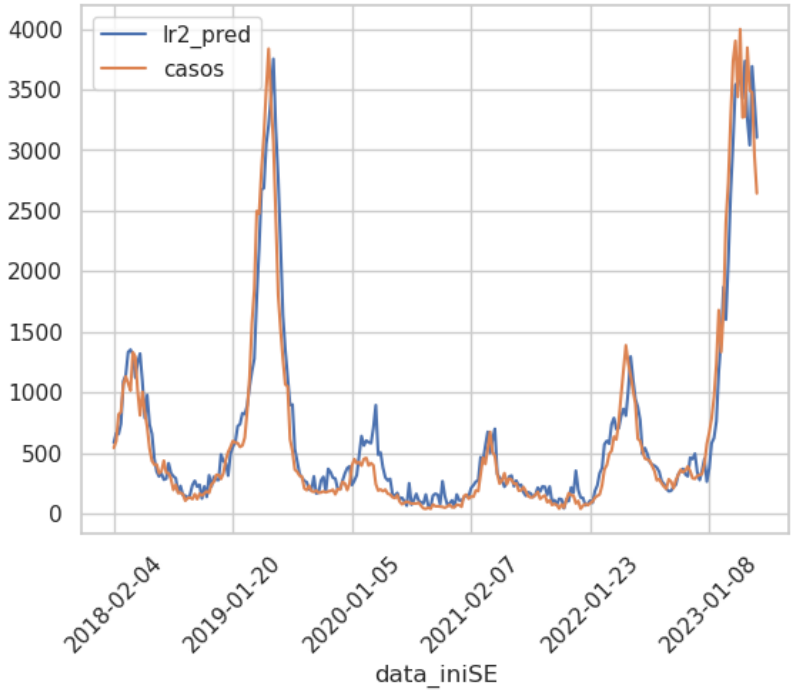

Figura 45 - Séries temporais de casos previstos e realizados em base treino e teste

Como é possível notar, o modelo parece ter performado supreendentemente bem e as previsões acompanharam a realidade mesmo nos principais picos, como os de 2019 e 2023.

Por fim, foi analisada a precisão das previsões em nível municipal e semestral, para dar uma dimensão de como o modelo desempenhou no local e no tempo. O mapa de calor da Figura 46 mostra o RMSE (Root Mean Squared Error) dividido pelo total de casos do município no mesmo período. No eixo x, são apresentados os semestres desde 2018.1 até 2023.1; e no eixo y, estão os municípios ordenados descendentemente de acordo com o total de casos de 2018 à 2023. A coloração está em escala logarítmica para melhorar a sensibilidade de acordo com as variações existentes e as partes vazias representam situações em que o total de casos foi zero, não permitindo a divisão.

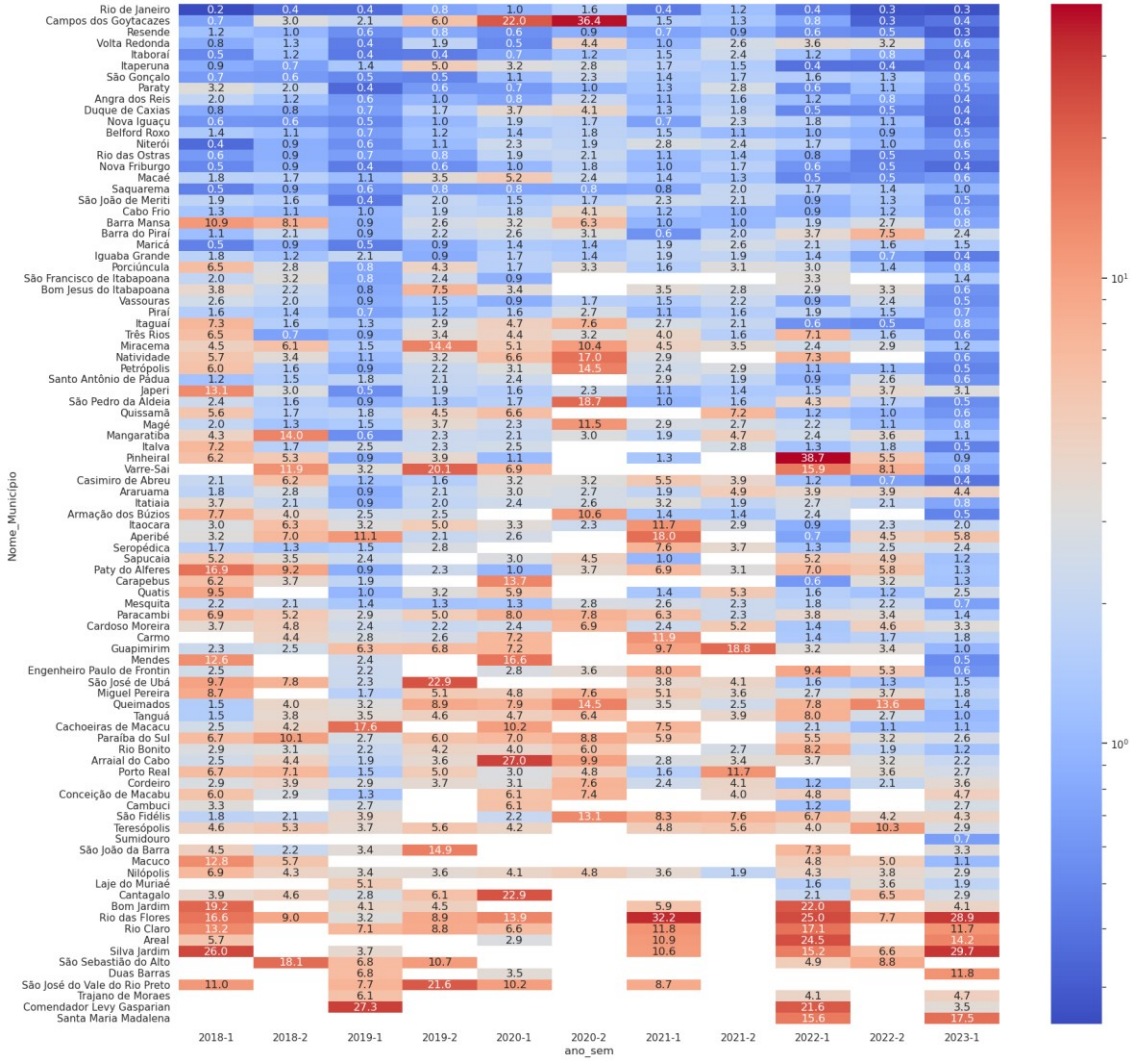

Figura 46 - Mapa de calor de RMSE relativo por município e semestre

Como é possível concluir, as predições se mostraram relativamente precisas para os municípios onde o número de casos é mais elevado, em especial o Rio de Janeiro. Quanto aos períodos de pior RMSE relativo em Campo dos Goytacazes, isso se deve ao fato do número de casos em 2020 ter sido baixo, elevando assim o erro relativo.

Além disso, vemos que as previsões também desempenharam satisfatoriamente no ano de 2023, especialmente se comparando aos anos anteriores. Tal constatação é relevante para reforçar a aplicabilidade do modelo, pois foi capaz de prever com qualidade a contaminação por dengue no mais recente surto de casos da doença no estado do Rio de Janeiro.

Tendo em vista todas as análises discutidas acima, podemos chegar à conclusão de que o modelo apresentou uma assertividade mais do que suficiente para seu propósito. Porém, essa não é a única entrega proporcionada por ele.

Uma das grandes vantagens de se ter usado um algoritmo como a regressão linear é pela interpretabilidade. A interpretabilidade de um modelo de aprendizado de máquina refere-se à facilidade com que usuário pode compreender as razões por trás das previsões ou decisões tomadas pelo modelo, ou seja, quais fatores estão contribuindo para as previsões e como as diferentes variáveis influenciam os resultados. Sobretudo na área da saúde, a interpretabilidade dos modelos é essencial e muitas vezes é tão importante quanto suas previsões.

Na Tabela 7, temos os coeficientes das variáveis do modelo e seus p-valores. Para que uma variável e seu coeficiente sejam considerados estatisticamente significativos, é necessário que o p-valor esteja abaixo de um limite de confiança. Para fins de simplificação, usamos como limite a faixa de 10% (intervalo de confiança de 90%) e separamos as variáveis em dois blocos: significativas e não significativas.

Avaliando as variáveis da Tabela 7, podemos fazer algumas conclusões.

Primeiro, concluímos que fatores sazonais, como mês e estação, são extremamente influentes no modelo. Isso porque não só são quase todas significativas – as exceções são os meses de outubro e novembro – como também tem coeficientes relevantes. Dentre os meses, os que tem maior coeficiente positivo são março, abril e maio, representando uma época de maior número de casos, seguidos pelos meses de junho e fevereiro. Já os meses do segundo semestre tem coeficientes mais baixos, quando não negativos, representando uma época de menor número de casos.

Em compensação, a estação de baixa, segunda variável mais significativa do modelo, tem um coeficiente de 5,256, o que significa que estar em época de baixa contribui com um piso de 5,256 casos. Comparando com os coeficientes dos meses da época de alta, notamos que a estação de baixa sozinha "contribui" com menos casos, porém, somando-o aos coeficientes meses que compõe esse período, resultaria em um piso de forma geral maior que o da época de alta.

Considerando a figura 25, que constatava que a estação de alta apresentava maior número de casos que a de baixa, o que explica então tais coeficientes? A resposta é a influência das demais variáveis. Ou seja, analisando sozinho o fator sazonalidade, entenderíamos que os meses e a estação de alta deveriam ter coeficientes mais altos que os demais. Entretanto, esses meses também são meses de temperatura mais alta, maior receptividade do clima, número de tweets e, naturalmente, de casos defasados de 2 semanas – todas variáveis significativas com coeficientes positivos. Levando em consideração todas essas variáveis em conjunto, podemos explicar, simplificadamente, que a alta de número de casos nesses meses do ano é contribuída parcialmente por cada um desses fatores e não apenas pelo fator sazonal em si. Em conjunto, justificam um número de casos mais elevado nesse período do ano.

Seguindo este raciocínio, evidenciou-se através desse modelo que a temperatura média atual é a variável climática de maior impacto, a julgar-se pela sua significância e seu coeficiente. Entretanto, a temperatura defasada de 21 dias, bem como todas as variáveis de precipitação e umidade não foram significativas para o modelo, além de apresentar coeficientes mais próximos de zero, denotando menor contribuição para o número de casos estimado. O que chamou a atenção também foi o coeficiente negativo da temperatura média de 2 meses antes.

Quanto às variáveis sociais e demográficas, podemos inferir algumas conclusões. Primeiramente e sem nenhuma surpresa, quanto maior a população, maior o número de casos. Em contrapartida, a densidade populacional influencia negativamente no número de casos, levando a entender que municípios menos povoados tenderiam a ter menor incidência da doença. Em contrapartida, pouco pode-se afirmar sobre a influência dos índices de desenvolvimento humano, pois todas estas variáveis obtiveram p-valor acima do corte de significância.

Ademais, vale apontar que o número de tweets sobre sintomas de dengue demonstrou ser um bom preditor para o número de casos de dengue e é uma variável importante para alarmar sobre o número de casos a serem confirmados de dengue, indicando aumento do número de casos.Para finalizar, há de se ressaltar que a principal variável do modelo foi o número de casos defasado de 2 semanas, confirmando um comportamento autorregressivo, já esperado, da série temporal de casos de dengue no estado do Rio de Janeiro. Essa variável não só teve o maior coeficiente como também teve a maior significância.

Significativas (p < 10%)

Variável	Coef.	p-valor
casos_lag14	**47,832**	0,000
estacao_baixa	**5,256**	0,000
receptivo_1	**5,144**	0,000
mes_4	**6,817**	0,000
mes_5	**6,385**	0,000
mes_3	**6,788**	0,000
mes_6	**5,662**	0,000
total	**1,796**	0,000
tweet	**1,355**	0,000
mes_2	**4,644**	0,000
tempmed	**1,070**	0,000
mes_8	**1,912**	0,000
mes_7	**1,641**	0,000
mes_9	**1,387**	0,000
tempmed_lag2	**-0,528**	0,015
densidade_pop	**-0,274**	0,027
mes_12	**-0,789**	0,064
Não Significativas (p > 10%)		
Variável	**Coeficiente**	**p-valor**
Domicilios_Saneamento	-0,187	0,118
mes_10	0,572	0,129
tempmed_lag21	0,318	0,131
mes_11	0,532	0,151
IDHM_Educacao	-0,239	0,152
precipitation_sum_ms_lag21	0,177	0,183
precipitation_sum_ms_lag2	0,199	0,200
umidmed_lag21	0,172	0,226
PIB_per_capita	0,119	0,309
precipitation_sum_ms_lag14	0,085	0,516
IDHM_Renda	-0,023	0,897

Tabela 7 - Coeficientes de variáveis e significâncias

5. Conclusão

O estudo apresentado neste trabalho ofereceu descobertas significativas acerca do comportamento dos índices de dengue no estado do Rio de Janeiro e os sobre os fatores que o influenciam. Através da análise de dados climáticos, socioeconômicos e de saúde, foi possível construir um modelo preditivo robusto que não apenas identifica tendências significativas, mas também possibilita a previsão do número de casos de dengue com um grau razoável de precisão.

A análise exploratória dos dados, em conjunto com o modelo previamente mencionado, resultou em uma série de descobertas significativas sobre as correlações e causas relacionadas à incidência de dengue, bem como sobre sua sazonalidade. Os conhecimentos obtidos em cada fase deste estudo se interligam, enriquecendo de forma substantiva a compreensão existente sobre a dengue, uma doença de relevância crítica e amplamente discutida no Brasil.

Primeiramente, observou-se que fatores climáticos, sobretudo a temperatura, desempenham um papel fundamental na proliferação do Aedes Aegypti, vetor da dengue. Em segundo lugar, o estudo ressaltou a importância de variáveis populacionais, como a densidade demográfica e indicadores de desenvolvimento humano.

Um aspecto crucial deste estudo foi a utilização de técnicas avançadas de análise de dados, incluindo diferentes algoritmos de aprendizado de máquina, para criar um modelo preditivo. Este modelo não só ajudou a identificar os fatores mais influentes na incidência de dengue, mas também provou ser uma ferramenta valiosa na previsão da propagação da doença. A aplicação de um modelo de regressão linear, em particular, mostrou-se eficaz, equilibrando precisão e interpretabilidade.

Por fim, este trabalho destaca a importância de uma abordagem multifatorial e interdisciplinar no estudo de doenças transmitidas por vetores como a dengue. Os resultados aqui apresentados podem ser valiosos para formuladores de políticas públicas, profissionais de saúde e pesquisadores, fornecendo uma base sólida para a tomada de decisões e para estratégias de prevenção e controle da dengue.

O modelo desenvolvido, embora eficaz, tem suas limitações e deve ser continuamente aprimorado com a integração de mais dados e a aplicação de técnicas analíticas avançadas. Um exemplo disso seria a eliminação de variáveis não significativas para o modelo, a fim de reduzir sua complexidade computacional. Além disso, seria de enorme importância expandir esse estudo para municípios de demais estados do país, providenciando uma rede mais ampla de monitoramento e prevenção da doença. Este trabalho abre caminho para futuras pesquisas que podem explorar ainda mais a relação entre fatores climáticos, socioeconômicos e a incidência de dengue, contribuindo para um controle mais efetivo e para a prevenção da doença no estado do Rio de Janeiro.

Referências

1. Atlas BR. Atlas de Desenvolvimento Humano no Brasil. Disponível em <http://www.atlasbrasil.org.br/>. Acesso em 5 dez. 2023.
2. Baturynska, I., Martinsen, K. Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. J Intell Manuf 32, 179–200 (2021). https://doi.org/10.1007/s10845-020-01567-0.
3. BEZERRA, Thiago de Matos; MATOS, Cintia Chagas. Dengue no Brasil: fatores socioambientais associados a prevalência de casos. Arquivos de Ciências da Saúde da UNIPAR. Umuarama, v.27, n.5, p.2685-2698, 2023.
4. BISHOP, Christopher M. Pattern Recognition and Machine Learning. 1 ed. Nova York: Springer, 2006.
5. BRANCO, Henrique. Overfitting e underfitting em machine learning. Disponível em <https://abracd.org/overfitting-e-underfitting-em-machine-learning/>. Acesso em 7 dez. 2023.
6. BREIMAN, Leo. Random Forests. Machine Learning. v. 45. p 5-32, 2001.
7. CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, San Francisco.
8. CRUZ, João Pedro. Bioestatística: quartil amostral. quartil amostral. 2022. Disponível em: https://sweet.ua.pt/pedrocruz/bioestatistica/ed-quartil-amostral.html. Acesso em: 25 nov. 2023.
9. DOS SANTOS, Alcione Miranda. Introdução à Estatística. Disponível em <https://docente.ifsc.edu.br/gianpaulo.medeiros/MaterialDidatico/M%C3%A9todos%20Est%C3%A1tisticos/estatistica%20aula%201.pdf>. Acesso em 8 dez. 2023.
10. Ferramenta que prevê a quantidade de mosquitos por região é lançada no Brasil. Revista Crescer. Rio de Janeiro, 28 mar. 2023. Disponível em <https://revistacrescer.globo.com/fique-por-dentro/noticia/2023/03/ferramenta-que-preve-a-quantidade-de-mosquitos-por-regiao-e-lancada-no-brasil.ghtml>. Acesso em 12 dez. 2023.
11. FREIRE, Sérgio Miranda. Bioestatística básica. Ed. do Autor: Rio de Janeiro, 2021. p. 559-584.
12. FREITAS, Thales. Entendendo as árvores de decisão em machine learning. Sigmoidal. São José dos Campos, 23 fev. 2022. Disponível em <https://sigmoidal.ai/entendendo-as-arvores-de-decisao-em-machine-learning/>. Acesso em 9 dez. 2023.
13. GRAVETTER, Frederick J; WALLNAU, Larry B. Statistics for the Behavioral Sciences. 10 ed. Boston: Cengage Learning, 2017.
14. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2 ed. Nova York: Springer, 2016.
15. HODSON, Timothy O. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. v. 15, p. 5481-5487, 2022.
16. HYNDMAN, Rob J; ATHANASOPOULOS, George. Forecasting: Principles and Practice. 2 ed. Melbourne: OTexts, 2018.
17. IBGE – Instituto Brasileiro de Geografia e Estatística. Sinopse do Censo Demográfico 2010: População residente, total, urbana total e urbana na sede municipal, em números absolutos e relativos, com indicação da área total e densidade demográfica, segundo as Unidades da Federação e os municípios. Disponível em <https://www.ibge.gov.br/censo2010/apps/sinopse/index.php?uf=33&dados=21>. Acesso em 25 nov. 2023.
18. Instituto Nacional de Meteorologia (INMET). Mapa de Estações. Disponível em <https://mapas.inmet.gov.br/>. Acesso em 5 dez. 2023.
19. LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey D. Mining of massive datasets. Cambridge: Cambridge University Press, 2014.
20. LESSA, Daniela; Oliveira, Claudio. Dengue: atividades de divulgação e pesquisas persistem no combate à doença. Portal Fiocruz. Rio de Janeiro, 20 fev. 2014. Disponível em <https://portal.fiocruz.br/noticia/dengue-atividades-de-divulgacao-e-pesquisas-persistem-no-combate-doenca>. Acesso em 22 nov. 2023.
21. Level up coding. Random forest regression. GitConnected, [2020]. Disponível em <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>. Acesso em 8 dez. 2023.
22. MACIEL, Fernanda. Como ler um boxplot. Disponível em <https://blog.proffernandamaciel.com.br/como-ler-um-boxplot/>. Acesso em 10 dez. 2023.
23. MURPHY, Kevin Patrick. Machine Learning: A Probabilistic Perspective. 1 ed. Cambridge: MIT Press, 2012.
24. NUNES, Priscila Conrado Guedes et al. 30 years of dengue fatal cases in Brazil: a laboratorial-based investigation of 1047 cases. BMC Infect Dis. v. 18, n. 346, p. 1-13, 2018.

25. NUNES, Priscila Conrado Guerra et al. 30 years of fatal dengue cases in Brazil: a review. BMC Public Health. v. 19, n. 329, p. 1-11, 2019.
26. O que é o IDHM. United Nations Development Programme (UNDP). Disponível em <https://www.undp.org/pt/brazil/o-que-%C3%A9-o-idhm#:~:text=O%20%C3%8Dndice%20de%20Desenvolvimento%20Humano,1%2C%20maior%20o%20desenvolvimento%20humano.>. Acesso em 7 dez. 2023.
27. O'NEIL, Cathy; SCHUTT, Rachel. Doing Data Science. 1 ed. Sebastopol: O'Reilly Media, 2013.
28. Open-Meteo. Free Weather API. Disponível em <https://open-meteo.com/>. Acesso em 7 dez. 2023.
29. PASQUINI, Patrícia. Brasil pode viver epidemia de dengue tipo 3 em 2024, diz consultor de braço da OMS. Folha de S. Paulo. São Paulo, 3 out. 2023. Disponível em <https://www1.folha.uol.com.br/equilibrioesaude/2023/10/brasil-pode-viver-epidemia-de-dengue-tipo-3-em-2024-diz-consultor-de-braco-da-oms.shtml>. Acesso em 10 dez. 2023.
30. ROSTER, Kirstin Ingrid Oliveira. Data science for epidemiology: a case study of dengue in Brazil. São Carlos. 2023. 119 p. Tese (Doutorado - Programa de Pós Graduação em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.
31. Sanchez-Gendriz, I., de Souza, G.F., de Andrade, I.G.M. et al. Data-driven computational intelligence applied to dengue outbreak forecasting: a case study at the scale of the city of Natal, RN-Brazil. Sci Rep 12, 6550 (2022). https://doi.org/10.1038/s41598-022-10512-5
32. Secretaria de Estado de Saúde de Mato Grosso. Informes de Dengue, Chikungunya e Zika. Disponível em <http://www.saude.mt.gov.br/dengue/pagina/419/chikungunya-sintomas#:~:text=De%20dois%20a%20dez%20dias,o%20v%C3%ADrus%20na%20corrente%20sangu%C3%ADnea.>. Acesso em 5 dez. 2023.
33. Secretaria de Vigilância em Saúde e Ambiente; Ministério da Saúde. Monitoramento dos casos de arboviroses até a semana epidemiológica 52 de 2022. Boletim Epidemiológico. Brasília, v. 54, n. 1, jan. 2023.
34. VALLE, Denise. Aedes e dengue: vetor e doença. Portal Fiocruz. Disponível em <https://www.ioc.fiocruz.br/dengue/textos/aedesvetoredoenca.html#:~:text=Para%20que%20a%20transmiss%C3%A3o%20da%20dengue%20aconte%C3%A7a%2C%20%C3%A9%20preciso%20que,vetor%20esteja%20infectado%20e%20infectivo&text=Ao%20mesmo%20tempo%20em%20que,maior%20volume%20poss%C3%ADvel%20de%20sangue>. Acesso em 22 nov. 2023.