ANÁLISE DE DADOS DA DENGUE NA REGIÃO OESTE PAULISTA

Thiago de Almeira Silva

Vanessa dos Anjos Borges

RESUMO: A dengue representa um dos principais desafios de saúde pública no Brasil, exigindo estratégias de monitoramento capazes de antecipar surtos e orientar medidas de prevenção, especialmente no Oeste Paulista. Este estudo analisou a incidência de casos de dengue nos 56 municípios do oeste paulista, investigando a relação entre variáveis epidemiológicas e climáticas e testando algoritmos de aprendizado de máquina para prever níveis de alerta. A metodologia adotada foi o CRISP-DM, contemplando compreensão, preparação, exploração e modelagem dos dados com apoio da ferramenta Orange Data Mining. A análise descritiva, por meio de Boxplots, distribuições de frequência e correlação, evidenciou padrões sazonais, com maior número de casos nos meses quentes e úmidos. As variáveis epidemiológicas (casos, Rt, p_inc100k, p_rt1) apresentaram maior associação com a dinâmica da doença, enquanto temperatura e umidade mostraram correlações consistentes, ainda que menos intensas. Na etapa de modelagem, compararam-se Árvore de Decisão e Floresta Aleatória, tendo como variável alvo o nível de alerta epidemiológico. A Árvore de Decisão alcançou acurácia de 93,0% e AUC de 0,819, oferecendo regras interpretáveis para identificação de fatores críticos. Já a Floresta Aleatória obteve acurácia de 93,7% e AUC de 0,972, com melhor desempenho na classificação das categorias intermediárias (amarelo e laranja). Esses resultados indicam o potencial da modelagem preditiva como suporte a sistemas de vigilância, permitindo a integração de dados em tempo real e a geração de dashboards interativos para gestores públicos. Como trabalhos futuros, propõe-se ampliar a base de dados e avaliar novos algoritmos, buscando maior robustez e aplicabilidade prática.

Palavras-chave: Data Mining. Análise epidemiológica. Saúde pública.

INTRODUÇÃO

A dengue, uma arbovirose transmitida pelo mosquito *Aedes aegypti* é um grave problema de saúde pública no Brasil, especialmente nas regiões urbanas e de clima tropical. Nos últimos

anos, o avanço da doença tem chamado atenção pelas altas taxas de contaminação, hospitalizações e óbitos. O Oeste Paulista exemplifica essa realidade.

Em 2025, Presidente Prudente foi a terceira cidade do país com o maior número de mortes por dengue, totalizando 33 óbitos, atrás das cidades de Goiânia (GO) e Brasília (DF) (G1, 2025). O cenário se agrava diante do aumento nos casos registrados em toda a região do interior de São Paulo, refletindo uma situação que exige atenção imediata e ações baseadas em evidências e dados epidemiológicos confiáveis.

A análise de dados sobre dengue tem se apresentado como uma ferramenta para compreender a dinâmica da doença e orientar políticas públicas mais eficazes. Conforme Ribeiro et al. (2022), a vigilância epidemiológica e o uso estratégico de informações quantitativas permitem o monitoramento mais preciso da disseminação do vírus e o direcionamento de recursos e esforços de prevenção em áreas mais vulneráveis. Além disso, estudos como o de Lopes e Souza (2023) apontam que a associação entre dados sociodemográficos e ambientais pode revelar padrões importantes de infecção e orientar medidas específicas em nível local.

Particularmente na região do oeste paulista, fatores como urbanização desordenada, variações climáticas, práticas inadequadas de saneamento e baixa eficácia nas campanhas de combate ao mosquito contribuem para a manutenção de ciclos epidêmicos (Oliveira et al., 2023). Frente a esse panorama, investir na coleta, organização e análise de dados com base científica pode subsidiar estratégias de controle mais eficientes e adaptadas à realidade local.

Sendo assim, este trabalho tem como objetivo analisar os dados da dengue nas cidades do oeste paulista, buscando identificar padrões de distribuição, fatores agravantes e possibilidades de intervenção.

PROCEDIMENTOS METODOLÓGICOS

Este estudo caracteriza-se como uma pesquisa aplicada, com abordagem quantitativa e caráter exploratório, voltada à análise de dados epidemiológicos referentes aos casos de dengue nas cidades do oeste do estado de São Paulo. A pesquisa buscou identificar padrões, tendências e fatores associados à incidência da doença, a fim de subsidiar estratégias de prevenção e controle mais eficazes.

Para a condução deste trabalho, adotou-se como referência metodológica o modelo CRISP-DM (*Cross Industry Standard Process for Data Mining*), amplamente utilizado em projetos de ciência de dados por sua estrutura sistemática e adaptabilidade a diferentes áreas do conhecimento (Niaksu, 2015). O modelo CRISP-DM organiza o processo analítico em seis etapas: compreensão do problema, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação.

Na fase de compreensão do problema, a partir de análise exploratória em trabalhos científicos, definiu-se como objetivo central compreender o comportamento da dengue na região do oeste paulista, que apresentou índices alarmantes de casos e mortalidade por dengue no país em 2025.

Na etapa de compreensão dos dados, foi realizada a análise de *datasets* provenientes de fontes públicas. Para o trabalho foi utilizado o sistema Info Dengue[1], um pipeline semi-automatizado que realiza a coleta e disponibilização de indicadores epidemiológicos da dengue e de outras arboviroses em nível municipal. O sistema foi implementado em 2015, desenvolvido por pesquisadores do Programa de Computação Científica (Fundação Oswaldo Cruz, RJ) e da Escola de Matemática Aplicada (Fundação Getúlio Vargas) com a forte colaboração da Secretaria Municipal de Saúde do Rio de Janeiro, o Observatório da Dengue/UFMG e pesquisadores da Universidade Federal do Paraná e da Universidade Estadual do Oeste do Paraná.

Segundo Oliveira et al. (2023), a organização e interpretação de dados confiáveis são essenciais para orientar políticas públicas e identificar fatores que contribuem para a persistência de surtos epidêmicos. Foram coletados os dados das 56 cidades que compõem o Oeste Paulista[2].

Para as fases de preparação, modelagem e avaliação, utilizou-se o software Orange Data Mining, uma ferramenta de código aberto voltada à análise de dados e aprendizado de máquina. O Orange permite a construção de fluxos de trabalho interativos por meio da conexão de componentes gráficos (*widgets*), facilitando a análise mesmo por usuários com pouca experiência em programação (Demsar et al., 2013). A escolha pela ferramenta se deu pela sua aplicabilidade

[1] Disponível em: https://info.dengue.mat.br/services/api

[2] Adamantina, Alfredo Marcondes, Álvares Machado, Anhumas, Caiabu, Caiuá, Dracena, Emilianópolis, Estrela do Norte, Euclides da Cunha Paulista, Flora Rica, Flórida Paulista, Iepê, Indiana, Inúbia Paulista, Irapuru, João Ramalho, Junqueirópolis, Lucélia, Marabá Paulista, Mariápolis, Martinópolis, Mirante do Paranapanema, Monte Castelo, Nantes, Narandiba, Nova Guataporanga, Osvaldo Cruz, Ouro Verde, Pacaembu, Panorama, Parapuã, Paulicéia, Piquerobi, Pirapozinho, Pracinha, Presidente Bernardes, Presidente Epitácio, Presidente Prudente, Presidente Venceslau, Rancharia, Regente Feijó, Ribeirão dos Índios, Rinópolis, Rosana, Sagres, Salmourão, Sandovalina, Santa Mercedes, Santo Anastácio, Santo Expedito, São João do Pau d'Alho, Taciba, Tarabai, Teodoro Sampaio, Tupi Paulista.

em estudos de saúde pública, conforme demonstrado por Matos, Souza e Reis (2019), que utilizaram o Orange para análises preditivas em contextos epidemiológicos.

RESULTADOS E DISCUSSÃO

O Quadro 1 apresenta o dicionário de dados extraídos do sistema InfoDengue, necessários para análise e definição de quais as variáveis relevantes para análises.

Quadro 1. Dicionário de dados do dataset do sistema InfoDengue

Campo	Descrição
data_ini_SE	Primeiro dia da semana epidemiológica (Domingo).
SE	Semana epidemiológica.
casos_est	Número estimado de casos por semana usando o modelo de *nowcasting*.
cases_est_min / cases_est_max	Intervalo de credibilidade de 95% do número estimado de casos.
casos	Número de casos notificados por semana (valores atualizados retrospectivamente todas as semanas).
p_rt1	Probabilidade de (Rt > 1). Para emitir alerta laranja, usa-se o critério p_rt1 > 0,95 por 3 semanas ou mais.
p_inc100k	Taxa de incidência estimada por 100.000 habitantes.
Localidade_id	Divisão submunicipal (implementada apenas no Rio de Janeiro).
nivel	Nível de alerta (1 = verde, 2 = amarelo, 3 = laranja, 4 = vermelho).
id	Índice numérico.
versao_modelo	Versão do modelo (uso interno).
Rt	Estimativa pontual do número reprodutivo de casos.
pop	População estimada (IBGE).
tempmin	Média das temperaturas mínimas diárias ao longo da semana.
tempmed	Média das temperaturas diárias ao longo da semana.
tempmax	Média das temperaturas máximas diárias ao longo da semana.
umidmin	Média da umidade relativa mínima diária do ar ao longo da semana.
umidmed	Média da umidade relativa diária do ar ao longo da semana.
umidmax	Média da umidade relativa máxima diária do ar ao longo da semana.
receptivo	Indica receptividade climática: 0 = desfavorável, 1 = favorável, 2 = favorável nesta semana e na passada, 3 = favorável por ≥ 3 semanas (ciclo completo de transmissão).
transmissao	Evidência de transmissão sustentada: 0 = nenhuma, 1 = possível, 2 = provável, 3 = altamente provável.
nivel_inc	Incidência: 0 = abaixo do limiar pré-epidemia, 1 = acima do pré-epidemia mas abaixo do epidêmico, 2 = acima do epidêmico.
notif_accum_year	Número acumulado de casos no ano.

Fonte: Elaborado pelos autores

A escolha das variáveis para análise da série temporal de casos de dengue na região do oeste paulista foi orientada por evidências presentes na literatura científica, que demonstram associação direta entre fatores climáticos, demográficos e indicadores epidemiológicos com a dinâmica de transmissão da doença.

Variáveis climáticas como temperatura mínima, média e máxima (tempmin, tempmed, tempmax), umidade relativa mínima, média e máxima (umidmin, umidmed, umidmax), além da variável "receptivo" (indicador de receptividade climática), foram incluídas com base em estudos que apontam seu papel determinante na proliferação do vetor *Aedes aegypti*. Um aumento da temperatura mínima, por exemplo, foi associado a um aumento significativo no número de casos de dengue em Londrina-PR, sendo identificado como um dos principais determinantes do comportamento sazonal da doença (Almeida et al., 2022). Além disso, a umidade do ar mostrou-se correlacionada de forma positiva com os casos de dengue em diferentes contextos urbanos (Almeida; Silva, 2017)

Variáveis epidemiológicas como o número de casos notificados (casos), casos estimados com *nowcasting*[3] (casos_est), a taxa de incidência estimada por 100 mil habitantes (p_inc100k), e o número reprodutivo estimado (Rt), fornecem base para análises de tendência, surtos e transmissibilidade. Esses indicadores são utilizados em modelagens preditivas, sendo o Rt um dos mais importantes parâmetros para prever expansão ou retração da transmissão (Chen; Moraga, 2024)

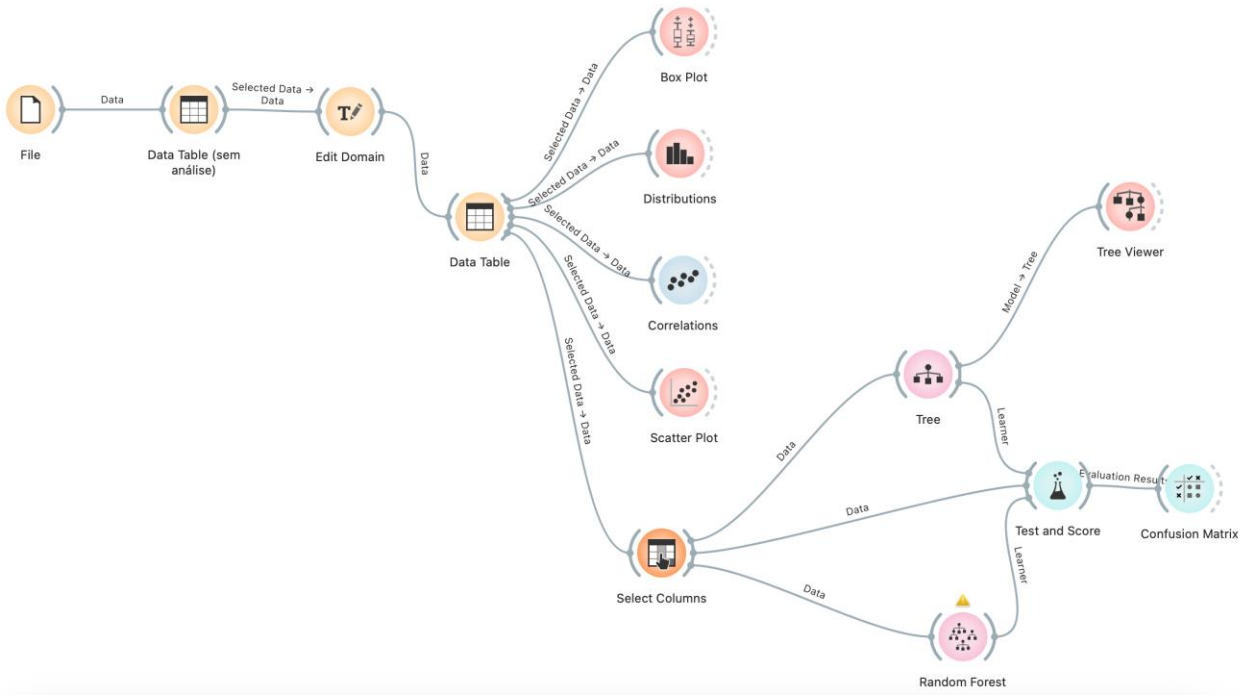
A inclusão da variável p_rt1 (probabilidade de Rt > 1) e do nível de alerta (nível) se justifica por sua aplicação direta em políticas públicas de prevenção, como observado em modelos de alerta precoce desenvolvidos para eventos de grande porte, como a Copa do Mundo de 2014 no Brasil (Lowe et al., 2014)

A população estimada (pop) também é um dado relevante para normalização dos dados e análise comparativa entre cidades de diferentes portes, como defendido em análises epidemiológicas multirregionais (Lira et al., 2021).

A partir das análises, iniciou-se a exploração e análise dos dados do *dataset* com a ferramenta Orange. A Figura 1 apresenta o fluxo de limpeza e preparação dos dados, no qual foram definidas as variáveis temporais, numéricas e categóricas.

[3] *Nowcasting* é uma prática de previsão de curtíssimo prazo, que se concentra em prever as condições atuais e futuras imediatas, seja em meteorologia (previsão de até 2 a 6 horas) ou em economia (previsão de um indicador econômico antes que os dados oficiais sejam publicados).

Figura 1: Fluxo de exploração e análise no Orange

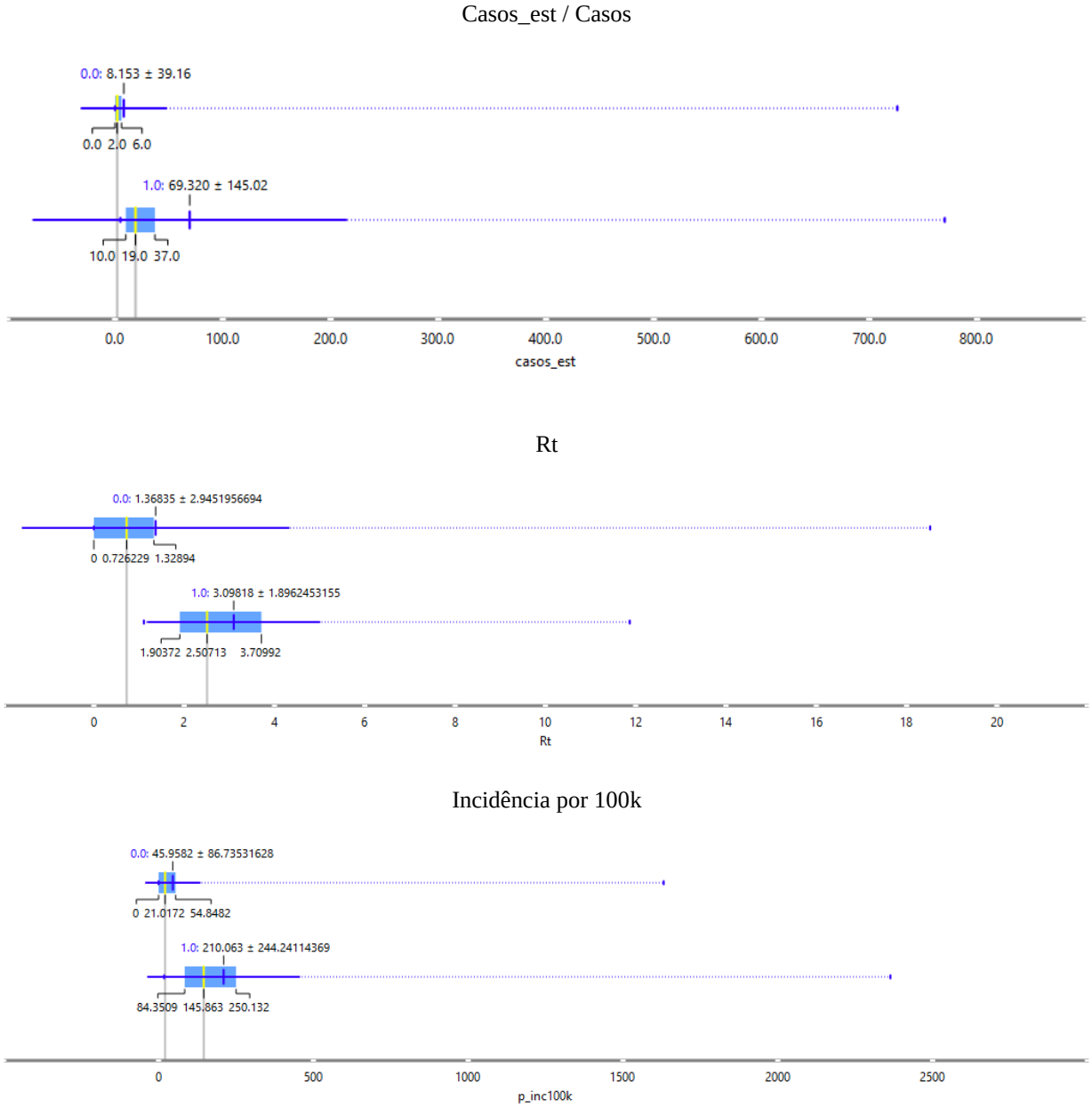


Fonte: Elaborado pelos autores

Foram realizadas a limpeza e preparação dos dados, no qual foram definidos os tipos das variáveis temporais, numéricas e categóricas (*widget* Select Columns). Na fase de exploração descritiva dos dados, foram aplicadas técnicas visuais que possibilitaram compreender o comportamento das variáveis epidemiológicas e climáticas ao longo do tempo. A Figura 2 apresenta os gráficos gerados pelo *widget* BoxPlot.

O BoxPlot é uma ferramenta estatística que permite visualizar de forma resumida a distribuição de uma variável numérica, destacando sua mediana, a variação entre os quartis e a presença de valores extremos (outliers), sendo útil para identificar assimetrias, dispersão dos dados e comparar diferentes grupos ou categorias, facilitando a detecção de padrões e comportamentos atípicos em um conjunto de informações (Krzywinski; Altman, 2014).

Figura 2: Gráficos BoxPlot das variáveis casos, casos_est, Rt e incidência por 100k



Fonte: Software *Orange Data Mining*, extraído pelos autores

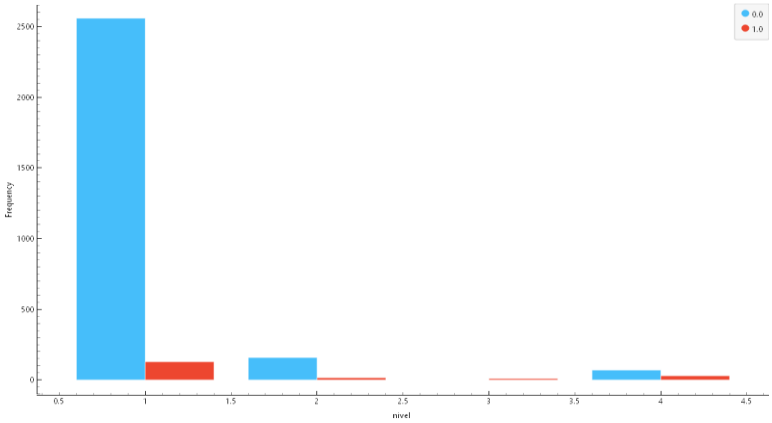
Por meio do Box Plot, verificou-se que as variáveis 'casos' e 'casos_est' apresentam uma distribuição com alta concentração de valores baixos (mediana em 2.0), indicando que a maioria das semanas epidemiológicas registra poucos casos. No entanto, a presença de valores máximos elevados (770.0) e o terceiro quartil em 7.0 demonstram a ocorrência de surtos e picos de casos, puxando a média para cima (11.88). A variável Rt (número reprodutivo) possui uma mediana de

0.79, sugerindo que, na maioria das semanas, a transmissão da doença está em declínio ou estável. Contudo, o valor máximo de 18.52 indica períodos de alta transmissibilidade. A taxa de incidência por 100 mil habitantes ('p_inc100k') segue um padrão similar, com mediana de 24.51 e picos de até 2363.51, refletindo a heterogeneidade na ocorrência da doença.

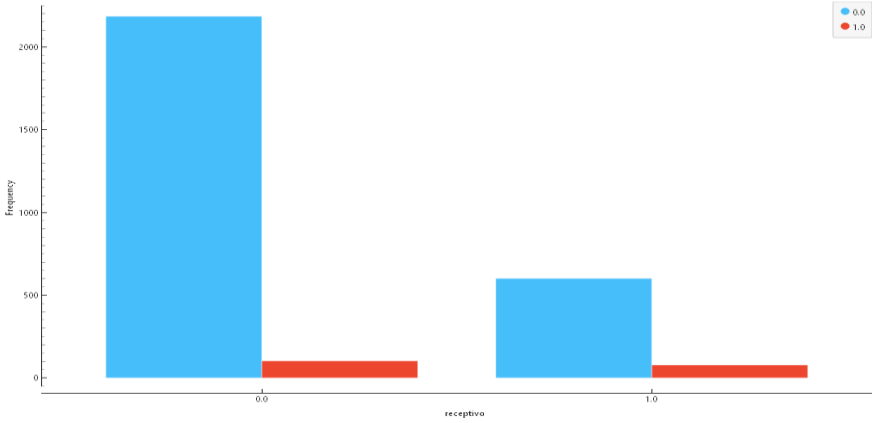
A Figura 3 apresenta os gráficos gerados pelo *widget* Distribution. O *widget* Distribution do Orange tem como propósito fornecer uma visualização da distribuição de frequências de uma variável, seja ela categórica ou numérica.

Figura 3: Gráficos do *widget* Distribution

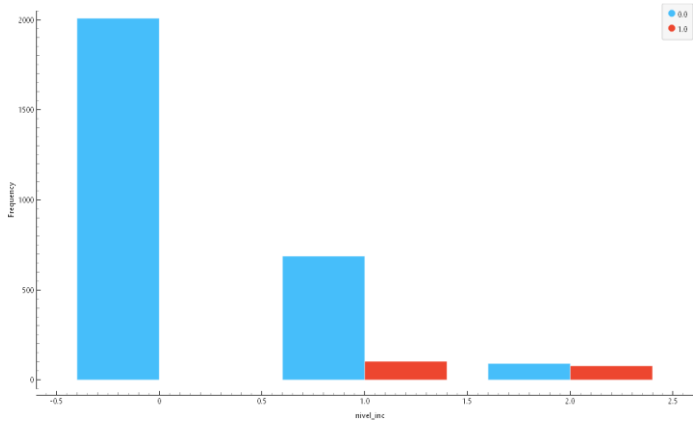
Nível



Receptivo



Nivel_inc

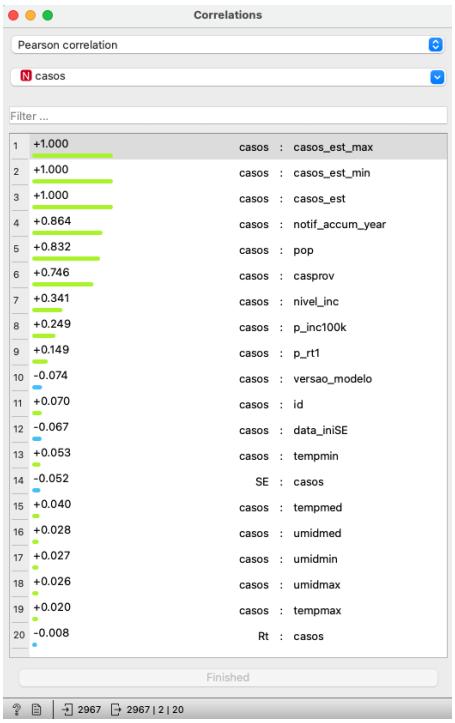


Fonte: Software *Orange Data Mining*, extraído pelos autores

A utilização do *widget* Distribution possibilitou analisar a frequência de ocorrência dos diferentes níveis de alerta e indicadores de transmissão. Observou-se que a maioria dos registros (2687) se encontra no 'nivel' 1 (verde), indicando que a doença está sob controle na maior parte do tempo. No entanto, há ocorrências nos níveis 2 (amarelo - 173 registros), 4 (vermelho - 98 registros) e 3 (laranja - 9 registros), confirmando a existência de períodos de alerta e surtos. Em relação à 'receptividade climática' ('receptivo'), a maioria das semanas (2288) foi classificada como desfavorável (0.0), enquanto 679 semanas foram favoráveis (1.0). Quanto ao 'nivel_inc' (incidência), 2008 registros estão abaixo do limiar pré-epidemia (0.0), 790 acima do pré-epidemia, mas abaixo do epidêmico (1.0), e 169 acima do epidêmico (2.0), reforçando a variabilidade da incidência ao longo do tempo.

Também foi utilizado o *widget* Correlation para analisar quais variáveis se associam mais fortemente com a variável casos. A Figura 4 apresenta o resultado apresentado pelo *widget*.

Figura 4: Resultado da análise apresentada pelo *widget* Correlation



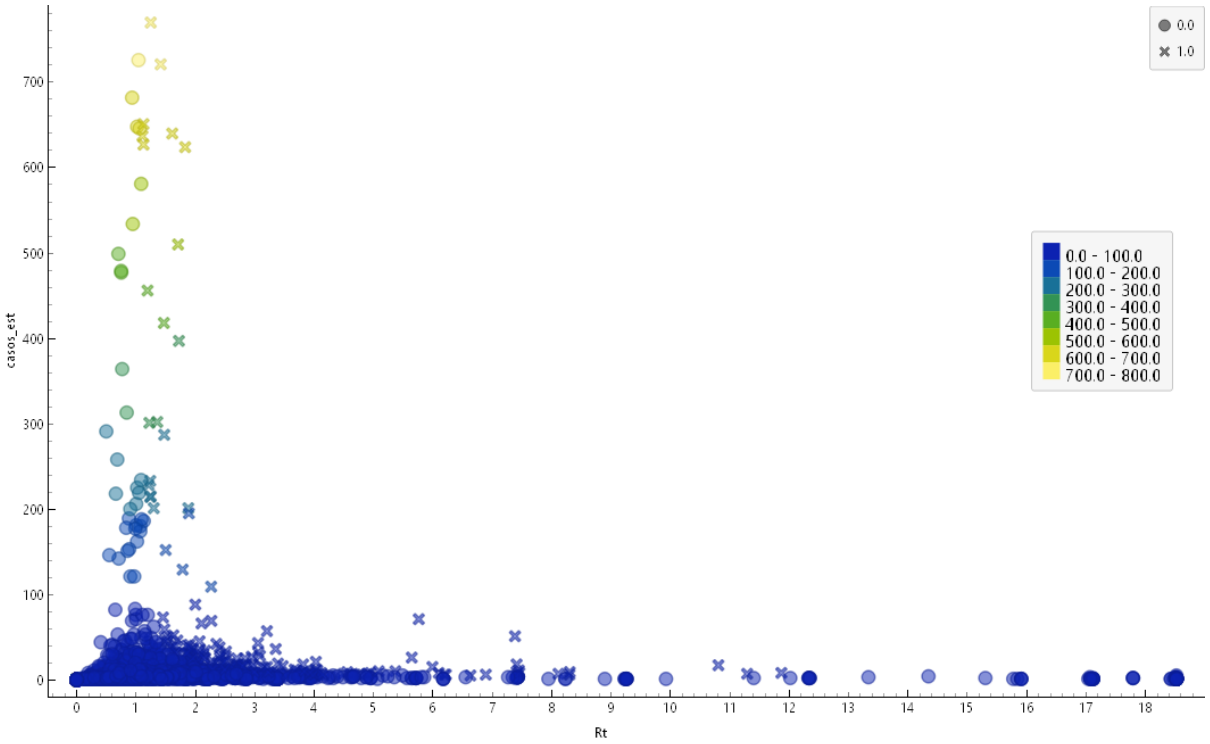
Fonte: Software *Orange Data Mining*, extraído pelos autores

Pelo conteúdo apresentado na Figura 4 é possível concluir que as variáveis 'casos', 'casos_est', 'casos_est_min' e 'casos_est_max' possuem correlação perfeita (1.0) entre si, o que é esperado, pois representam a mesma métrica de casos. Variáveis como 'notif_accum_year' (0.86), 'pop' (0.83) e 'casprov' (0.75) apresentam forte correlação positiva com o número de casos, indicando que municípios com maior população e maior número de casos prováveis tendem a ter mais casos notificados.

Indicadores epidemiológicos como 'nivel_inc' (0.34), 'nivel' (0.32) e 'transmissao' (0.27) também mostram correlação positiva moderada, sugerindo que níveis mais altos de incidência e alerta estão associados a um maior número de casos. Variáveis climáticas como 'tempmin' (0.053), 'tempmed' (0.040), 'umidmed' (0.028), 'umidmin' (0.027) e 'umidmax' (0.026) apresentam correlações positivas, mas muito fracas, com o número de casos, indicando uma influência limitada ou não linear direta. Curiosamente, 'Rt' (número reprodutivo) mostra uma correlação negativa muito fraca (-0.007), e 'SE' (semana epidemiológica) uma correlação negativa fraca (-0.051), o que pode indicar complexidade nas relações temporais e de transmissão.

O *widget* Scartterplot foi utilizado para analisar a relação entre estimativa pontual do número reprodutivo de casos (Rt) e o número estimado de casos por semana usando o modelo de *nowcasting* (valores atualizados retrospectivamente a cada semana - casos_est). A Figura 5 apresenta o resultado apresentado pelo *widget*.

Figura 5: Resultado da análise apresentada pelo *widget* Scartterplot



Fonte: Software *Orange Data Mining*, extraído pelos autores

É possível observar na Figura 5 que a correlação entre 'Rt' (número reprodutivo) e 'casos_est' (casos estimados) é extremamente baixa e negativa (-0.0077). Isso sugere que não há uma relação linear direta clara entre a estimativa pontual do número reprodutivo e o número de casos estimados por semana. Embora o Rt seja um indicador importante da transmissibilidade da doença, sua relação com o volume absoluto de casos pode ser complexa e influenciada por outros fatores não lineares ou defasagens temporais, que não são capturados por uma análise de correlação linear simples.

Para garantir a consistência da base e a adequação dos dados aos algoritmos aplicados nas etapas seguintes, possibilitando análises descritivas, identificação de padrões sazonais e

modelagem preditiva do risco de dengue foi realizada a padronização dos atributos contínuos para reduzir diferenças de escala, utilizando o *widget* Continuize.

Optou-se pela normalização das variáveis numéricas utilizando o método de padronização *z-score*, que transforma os valores de cada atributo para média zero e desvio padrão igual a um. Essa escolha deve-se ao fato de que os atributos presentes no *dataset*, como número de casos notificados, taxa de incidência, temperatura e umidade, apresentam escalas bastante distintas. Sem a padronização, variáveis com valores absolutos maiores poderiam exercer maior influência sobre os algoritmos de análise. Dessa forma, o uso do *z-score* garantiu comparabilidade entre os atributos, permitindo identificar padrões e relações de forma mais equilibrada e confiável.

A partir da análise exploratória, foram selecionadas as variáveis mais relevantes para a modelagem preditiva. As variáveis epidemiológicas (casos, p_inc100k, Rt e p_rt1) mostraram-se diretamente associadas à incidência da doença. As variáveis categóricas (nivel_inc, receptivo, transmissao) forneceram informações adicionais sobre condições de risco, enquanto as variáveis climáticas (tempmin, tempmed, tempmax, umidmin, umidmed, umidmax) foram mantidas na modelagem pelo papel reconhecido da temperatura e da umidade na dinâmica do vetor. A variável demográfica pop foi incluída para permitir comparações entre cidades de diferentes portes populacionais. Variáveis redundantes, como casos_est, casos_est_min e casos_est_max, foram desconsideradas devido à alta correlação com casos, evitando sobreposição de informação no modelo.

O Quadro 2 apresenta a relação entre as variáveis exploradas na etapa descritiva, os principais achados que sustentaram sua relevância e a justificativa para sua inclusão nos modelos preditivos.

Quadro 2 – Relação entre variáveis, análise descritiva e justificativa para uso nos modelos preditivos

Variável	**Evidência na Análise Descritiva**	**Justificativa para uso na Modelagem**
casos	Boxplot mostrou concentração de valores baixos com picos elevados (surtos)	Indicador direto da ocorrência da doença; base para prever níveis de alerta.
p_inc100k	Boxplot revelou grande variação, incluindo valores extremos	Normaliza casos por população, permitindo comparação entre municípios.
Rt	Boxplot indicou mediana <1 mas com picos elevados; Scatterplot mostrou relação complexa com casos	Indicador chave de transmissibilidade; ajuda a antecipar surtos.
p_rt1	Relacionado a políticas de alerta (critério p_rt1 > 0,95 por 3 semanas)	Captura a probabilidade de expansão da transmissão; relevante para classificação de risco.

nivel_inc	Distribution mostrou distribuição entre limiar pré-epidêmico e epidêmico	Indicador epidemiológico de transição entre fases da epidemia.
receptivo	Distribution mostrou maioria em condição desfavorável, mas também semanas favoráveis	Expressa condições climáticas favoráveis ao vetor; útil para previsão de risco.
transmissao	Correlação positiva moderada com casos	Mede evidência de transmissão sustentada; importante para classificar risco.
pop	Correlação forte com casos	Permite ajustar comparações entre cidades de diferentes portes populacionais.
tempmin, tempmed, tempmax	Correlações fracas, mas com base teórica em estudos anteriores	Determinantes da sazonalidade da dengue; influência não linear esperada.
umidmin, umidmed, umidmax	Correlações fracas, mas sustentadas pela literatura	Umidade afeta a sobrevivência do vetor; variável relevante mesmo com correlação baixa.
notif_accum_year	Correlação forte com casos	Indica carga acumulada da doença no ano; sinaliza intensidade epidêmica.

Fonte: Elaborado pelos autores

Sendo assim, buscou-se avançar para a etapa de análise preditiva, com o objetivo de identificar padrões e avaliar a capacidade dos atributos climáticos e epidemiológicos em explicar os níveis de risco da dengue. Para isso, foram empregados os *widgets* Tree e Random Forest da ferramenta Orange, que possibilitam a construção de modelos supervisionados de classificação. Esses algoritmos foram selecionados por sua capacidade de lidar tanto com variáveis numéricas quanto categóricas e por fornecerem resultados que permitem, simultaneamente, interpretação das regras de decisão e maior robustez preditiva.

O algoritmo de Árvore de Decisão (Tree) é um método de aprendizado supervisionado que busca construir um modelo preditivo a partir da divisão recursiva dos dados em subconjuntos homogêneos. Cada divisão (nó) é realizada com base em uma variável que melhor separa os registros segundo a classe de interesse, formando uma estrutura hierárquica de regras do tipo “se... então...” (Tangirala, 2020).

No contexto da análise da dengue, a árvore de decisão permite identificar relações diretas entre variáveis climáticas (como temperatura e umidade) e epidemiológicas (como Rt e casos estimados) com os níveis de alerta (nivel). O principal propósito da utilização desse *widget* é possibilitar a interpretação transparente e intuitiva dos fatores que mais influenciam na classificação de risco, facilitando a explicação dos resultados para tomadores de decisão e gestores de saúde.

A visualização do modelo de Árvore de Decisão no Tree mostra como as variáveis epidemiológicas e climáticas foram utilizadas para classificar os níveis de alerta (nivel). O nó raiz da árvore confirma a importância de indicadores epidemiológicos diretos, como número de casos e Rt, que aparecem como critérios iniciais de divisão. Isso está em linha com os Boxplots e distribuições apresentados na análise descritiva, que já evidenciavam grande variabilidade no número de casos e a relevância da transmissibilidade.

À medida que a árvore se ramifica, observa-se que variáveis como probabilidade de Rt>1 (p_rt1), incidência por 100 mil habitantes (p_inc100k) e o nível de incidência (nivel_inc) são usadas para refinar a classificação. Esses atributos funcionam como limiares que discriminam as situações de baixo risco (verde e amarelo) das de maior risco (laranja e vermelho). A presença de variáveis climáticas, como temperatura média e umidade relativa, em ramos mais profundos da árvore reforça que elas não são determinantes principais, mas complementam a previsão em cenários limítrofes, corroborando a análise descritiva que apontava correlações fracas, mas teoricamente justificadas.

Outro aspecto relevante é que os nós terminais da árvore mostram maior concentração de classificações corretas para nível verde, reflexo do desbalanceamento da base de dados. Já os níveis laranja e vermelho aparecem em ramos mais específicos e profundos, indicando que o modelo consegue identificá-los, mas exige combinações mais restritas de variáveis (ex.: altos valores de Rt combinados com receptividade climática e incidência elevada).

Para a construção do modelo de árvore de decisão, foram definidos parâmetros que visam equilibrar precisão e interpretabilidade. Estabeleceu-se que cada folha deveria conter no mínimo dois exemplos (*Min. number of instances in leaves* = 2) e que não seriam realizadas divisões em subconjuntos com menos de cinco registros (*Do not split subsets smaller than* = 5), a fim de evitar regras baseadas em casos isolados. A profundidade máxima da árvore foi limitada a 10 níveis, de modo a garantir maior clareza na visualização das regras geradas, prevenindo a criação de modelos excessivamente complexos e sobreajustados. Além disso, foi adotado o critério de parada quando 95% das instâncias de um nó pertenciam a uma mesma classe (*Stop when majority reaches* [%] = 95), o que contribui para reduzir ramificações desnecessárias. Essa configuração permitiu gerar uma árvore suficientemente detalhada para identificar padrões relevantes, mas ainda interpretável e útil para aplicação prática em vigilância epidemiológica.

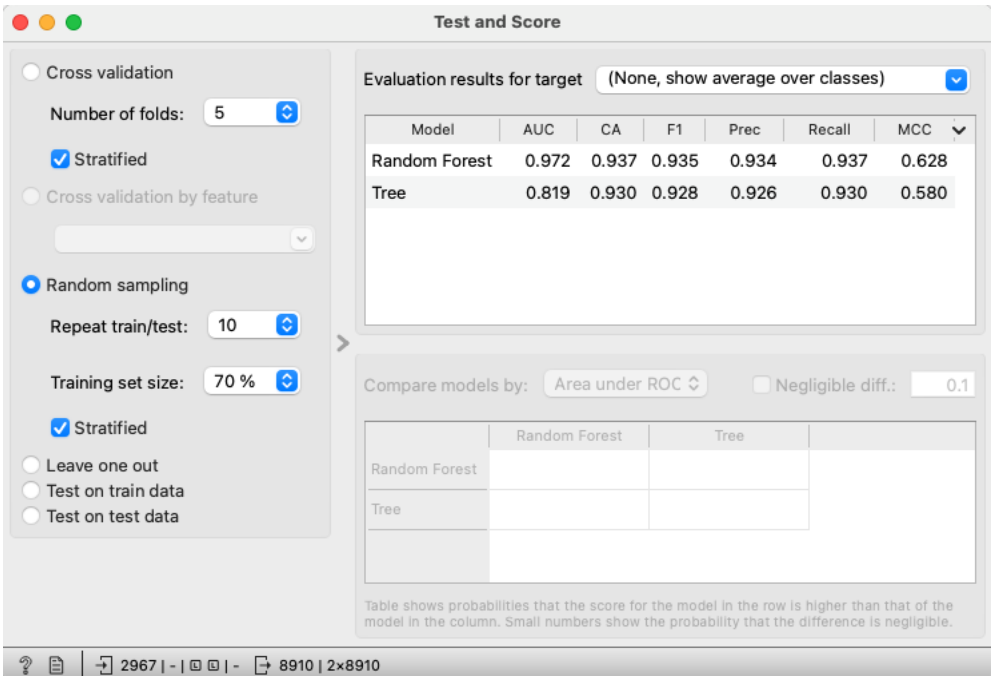
O algoritmo de Floresta Aleatória (Random Florest) é uma extensão das árvores de decisão, baseada em um conjunto (*ensemble*) de múltiplas árvores construídas a partir de amostras aleatórias do conjunto de dados. Cada árvore contribui com um "voto", e a classificação final é definida pela maioria. Esse método reduz a variância e o risco de sobreajuste (overfitting) presentes em uma única árvore, aumentando a robustez e a capacidade de generalização do modelo (Kulkarni; Sinha; Petare, 2016).

No caso da dengue, a floresta aleatória permite avaliar com maior precisão a relação entre múltiplos atributos (climáticos, epidemiológicos e demográficos) e os níveis de alerta, além de gerar métricas de importância das variáveis, apontando quais atributos mais contribuem para a previsão do risco.

No modelo de Floresta Aleatória, foram ajustados parâmetros para aumentar a robustez e lidar com o desbalanceamento das classes. O número de árvores foi definido em 100, valor que garante estabilidade nos resultados sem comprometer o desempenho computacional. Foi ativada a opção *Balance class distribution*, de modo a atribuir maior peso às classes minoritárias (nivel 3 e 4), o que favorece a capacidade do modelo em reconhecer situações críticas de risco epidemiológico. Embora essa configuração possa reduzir a acurácia global — como indicado pelo aviso do software — ela é metodologicamente justificável, pois privilegia a identificação de alertas relevantes para a saúde pública, mesmo quando esses representam uma minoria no conjunto de dados. Além disso, manteve-se a restrição de não dividir subconjuntos menores que cinco registros, a fim de evitar a geração de regras baseadas em casos isolados.

Para a avaliação dos modelos, foi utilizado o widget Test & Score, adotando o método de amostragem aleatória estratificada, com divisão de 70% dos dados para treino e 30% para teste, repetido em 10 execuções. Essa estratégia permitiu avaliar a robustez dos modelos frente ao desbalanceamento das classes. Foram analisadas métricas como acurácia, F1-score e AUC, possibilitando comparar o desempenho da Árvore de Decisão e da Floresta Aleatória. A Figura 6 apresenta o resultado dessa avaliação.

Figura 6: Resultado obtido no *widget* Test and Score

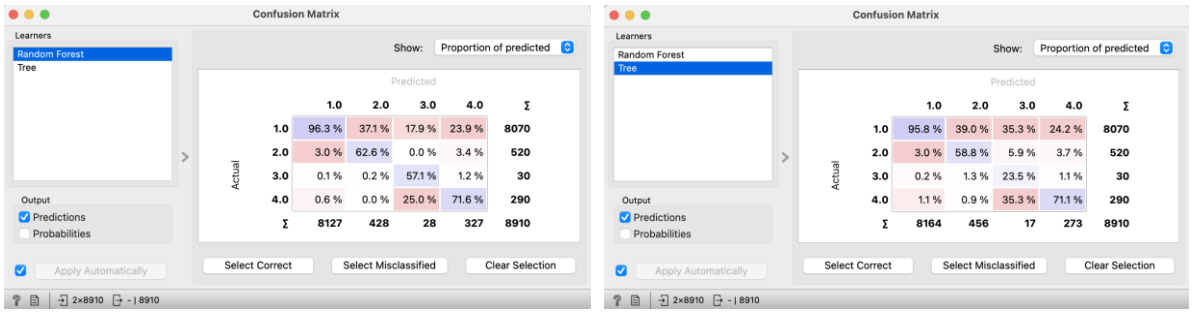


Fonte: Software *Orange Data Mining*, extraído pelos autores

A inclusão das variáveis identificadas na análise descritiva, contemplando tanto indicadores epidemiológicos quanto climáticos e demográficos, resultou em um desempenho satisfatório dos modelos avaliados. A Árvore de Decisão apresentou métricas consistentes, com acurácia de 93,0% (CA = 0,930), F1-score de 0,928 e AUC de 0,819, confirmando sua utilidade como ferramenta interpretável para compreensão das regras de classificação. No entanto, a Floresta Aleatória superou esses resultados, alcançando AUC de 0,972, acurácia de 93,7% (CA = 0,937) e F1-score de 0,935, além de apresentar maior correlação de Matthews (MCC = 0,628 contra 0,580 da Árvore). Esses resultados demonstram que, ao integrar múltiplos atributos epidemiológicos e ambientais, a Floresta Aleatória consegue capturar melhor as relações complexas entre variáveis, oferecendo maior capacidade discriminativa entre os níveis de alerta da dengue, sem perda de generalização.

A Figura 7 apresenta as Matrizes de Confusão que detalham as previsões realizadas pelos modelos.

Figura 7. Matrizes de Confusão dos modelos Árvore de Decisão e Árvore Aleatória



Fonte: Software *Orange Data Mining*, extraído pelos autores

A análise das matrizes de confusão evidencia o desempenho distinto entre os dois modelos utilizados. Ambos apresentaram alto índice de acerto para a classe Verde (nível 1), com a Floresta Aleatória atingindo 96,3% de classificações corretas e a Árvore de Decisão 95,8%, demonstrando que os cenários de baixo risco epidemiológico são bem identificados por ambos os algoritmos.

No entanto, nas categorias intermediárias, surgem diferenças relevantes: para o nível Amarelo (classe 2), a Floresta Aleatória obteve 62,6% de acerto, contra 58,8% da Árvore, mostrando maior capacidade de discriminação nessa transição de risco. O nível Laranja (classe 3) foi o mais problemático em ambos os modelos, mas a Floresta Aleatória obteve desempenho superior, com 57,1% de acerto, frente a apenas 23,5% da Árvore, que frequentemente confundiu casos Laranja com as categorias Vermelho e Amarelo.

Já para o nível Vermelho (classe 4), ambos os modelos apresentaram desempenho satisfatório, com 71,6% de acerto na Floresta Aleatória e 71,1% na Árvore, ainda que houvesse confusão significativa com a classe Laranja.

A Floresta Aleatória demonstrou maior robustez, especialmente nas classes intermediárias (Amarelo e Laranja), onde as fronteiras epidemiológicas são mais sutis e os indicadores apresentam maior sobreposição. A Árvore de Decisão, embora mais simples e interpretável, teve maior tendência a confundir essas classes, reforçando a vantagem de métodos ensemble quando o objetivo é obter previsões mais precisas para cenários complexos de alerta epidemiológico.

CONSIDERAÇÕES FINAIS

Ao compreender a realidade epidemiológica regional, espera-se contribuir para o aprimoramento das políticas públicas de saúde e para a redução dos impactos da doença sobre a população.

Ao longo deste estudo, foram explorados e analisados os dados da dengue na região Oeste Paulista, buscando encontrar os padrões e identificar complexidades. Utilizando Orange Data Mining foi possível compreender mais profundamente esse cenário epidemiológico.

A partir das análises realizada é possível concluir que a região é suscetível a picos e surtos significativos de casos de dengue. Isso ressalta a importância de uma vigilância contínua e da capacidade de resposta rápida das autoridades de saúde. Fatores como o número acumulado de notificações anuais, a população e os casos prováveis têm uma forte relação com a ocorrência da doença, enquanto as variáveis climáticas, embora relevantes, mostraram uma correlação mais sutil e complexa, sugerindo que sua influência pode ser mais indireta ou atuar em conjunto com outros fatores.

No campo da modelagem preditiva, os algoritmos de Árvore de Decisão e, principalmente, o Random Forest, se mostraram ferramentas promissoras para prever os níveis de alerta de dengue.

Apesar dos resultados promissores, o estudo apresenta limitações, como o desbalanceamento de classes (níveis laranja e vermelho pouco representados) e a ausência de variáveis entomológicas ou socioeconômicas. Estudos futuros podem explorar técnicas adicionais, como XGBoost e redes neurais, além da integração de dados entomológicos e de mobilidade populacional, ampliando a capacidade preditiva dos modelos.

Os resultados deste trabalho são um convite à ação. A Ciência de Dados não é apenas sobre números e algoritmos; é sobre capacitar gestores e profissionais de saúde com informações precisas para tomar decisões mais eficazes. Ao compreender melhor os fatores que impulsionam a dengue, pode-se direcionar recursos de forma mais inteligente, fortalecer campanhas de conscientização e implementar medidas preventivas mais assertivas, protegendo assim a saúde e o bem-estar de nossa população.

A partir dessas conclusões, abre-se espaço para o desenvolvimento de ferramentas computacionais que auxiliem gestores de saúde pública na tomada de decisão. Um software

baseado em modelos de aprendizado de máquina, como a Floresta Aleatória, poderia ser implementado para analisar automaticamente séries temporais de casos de dengue e variáveis climáticas, classificando semanalmente os municípios em níveis de alerta (verde, amarelo, laranja e vermelho). Essa solução poderia integrar dados de diferentes fontes — como sistemas epidemiológicos oficiais e bases meteorológicas — e gerar relatórios visuais e preditivos acessíveis em painéis interativos. Com isso, autoridades sanitárias teriam acesso a um sistema de apoio à decisão em tempo real, capaz de antecipar surtos, otimizar a alocação de recursos e orientar campanhas preventivas de forma mais direcionada.

REFERÊNCIAS

ALMEIDA, Daniela Sanches et al. Estudo da relação entre variáveis meteorológicas e ocorrência de casos de dengue em Londrina–PR. **Revista Brasileira de Geografia Física**, v. 14, n. 7, p. 3857-3866, 2021. Disponível em: https://periodicos.ufpe.br/revistas/index.php/rbgfe/article/view/250061. Acesso em: 22 ago. 2025.

ALMEIDA, Caio Américo Pereira; SILVA, Richarde Marques. Modelagem espacial dos casos de dengue e variáveis socioambientais em João Pessoa, Cabedelo e Bayeux, Paraíba. **Revista Brasileira de Geografia Física**, v. 10, n. 05, p. 1455-1470, 2017. Disponível em: https://periodicos.ufpe.br/revistas/index.php/rbgfe/article/view/234113. Acesso em: 22 ago. 2025.

CHEN, Xiang; MORAGA, Paula. Assessing dengue forecasting methods: a comparative study of statistical models and machine learning techniques in Rio de Janeiro, Brazil. **Tropical medicine and health**, v. 53, n. 1, p. 52, 2025. Disponível em: https://link.springer.com/article/10.1186/s41182-025-00723-7. Acesso em: 22 ago. 2025.

DEMŠAR, J. et al. Orange: Data Mining Toolbox in Python. **Journal of Machine Learning Research**, v. 14, p. 2349–2353, 2013. Disponível em: https://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf. Acesso em: 21 ago. 2025.

G1. Presidente Prudente é a terceira cidade do país com mais mortes por dengue em 2025; veja ranking. G1, 24 abr. 2025. Disponível em: https://g1.globo.com/sp/presidente-prudente-regiao/noticia/2025/04/24/presidente-prudente-e-a-terceira-cidade-do-pais-com-mais-mortes-por-dengue-em-2025-veja-ranking.ghtml. Acesso em: 21 ago. 2025.

KRZYWINSKI, Martin; ALTMAN, Naomi. Visualizing samples with box plots: use box plots to illustrate the spread and differences of samples. **Nature Methods**, v. 11, n. 2, p. 119-121, 2014. Disponível em: https://go.gale.com/ps/i.do?id=GALE%7CA361242515&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=15487091&p=HRCA&sw=w&userGroupName=anon~e3612f44&aty=open-web-entry. Acesso em: 25 ago. 2025.

KULKARNI, Vrushali Y.; SINHA, Pradeep K.; PETARE, Manisha C. Weighted hybrid decision tree model for random forest classifier**. Journal of The Institution of Engineers (India): Series B**, v. 97, p. 209-217, 2016. Disponível em: https://link.springer.com/article/10.1007/s40031-014-0176-y. Acesso em: 22 ago. 2025.

LIRA, Larine Ferreira et al. Incidência da dengue no Brasil: análise comparativa entre São Paulo e Alagoas Dengue incidence in Brazil: comparative analysis between São Paulo and Alagoas. **Brazilian Journal of Health Review**, v. 4, n. 6, p. 24410-24426, 2021. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BJHR/article/view/39352. Acesso em: 22 ago. 2025.

LOPES, A. A.; SOUZA, R. A. M. Análise epidemiológica e controle da dengue: o uso de geotecnologias em áreas urbanas. **Brazilian Journal of Innovation and Health Sciences**, v. 3, n. 2, p. 59-72, 2023. Disponível em: https://bjihs.emnuvens.com.br/bjihs/article/view/1781/2011. Acesso em: 21 ago. 2025.

LOWE, Rachel et al. Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. **The Lancet infectious diseases**, v. 14, n. 7, p. 619-626, 2014. Disponível em: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(14)70781-9/abstract?appunica=. Acesso em: 22 ago. 2025.

MATOS, Fernanda Fernandes; SOUZA, Renato Rocha; REIS, Zilma Silveira Nogueira. Análise de dados de saúde: mineração de texto com a utilização do Orange Canvas para exploração da informação. **Encontro Nacional de Pesquisa e Pós-graduação em Ciência da Informação**, 2019. Disponível em: https://repositorio.ufmg.br/handle/1843/56724. Acesso em: 21 ago. 2025.

NIAKSU, Olegas. CRISP data mining methodology extension for medical domain. Baltic Journal of Modern Computing, v. 3, n. 2, p. 92, 2015. Disponível em: https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/3_2_2_Niaksu.pdf. Acesso em: 21 ago. 2025.

OLIVEIRA, M. F. et al. Dengue e saúde pública: desafios e perspectivas para o enfrentamento no Brasil. **SciELO Preprints**, 2023. Disponível em: https://preprints.scielo.org/index.php/scielo/preprint/view/8333/15565. Acesso em: 21 ago. 2025.

RIBEIRO, A. C. F. et al. Aplicação de métodos estatísticos para análise espacial de epidemias: uma revisão narrativa. **Acta Paulista de Enfermagem**, v. 35, eAPE03271, 2022. Disponível em: https://www.scielo.br/j/ape/a/krgPGsgxLr8VSzkBhm9Qw9q/?format=html&lang=pt. Acesso em: 21 ago. 2025.

TANGIRALA, Suryakanthi. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. **International Journal of Advanced Computer Science and Applications**, v. 11, n. 2, p. 612-619, 2020. Disponível em: https://thesai.org/Publications/ViewPaper?Volume=11&Issue=2&Code=IJACSA&SerialNo=77. Acesso em: 22 ago. 2025.