UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ - UNIOESTE

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

Aplicação de modelos estatísticos e de aprendizagem de máquina na predição de casos de dengue em Cascavel - PR

Trabalho de Conclusão de Curso

Angelo José Orssatto

Cascavel–PR

2021

UNIVERSIDADE ESTADUAL DO OESTE DO PARANÁ - UNIOESTE

CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

Colegiado de Ciência da Computação

Curso de Bacharelado em Ciência da Computação

Angelo José Orssatto

Aplicação de modelos estatísticos e de aprendizagem de máquina na predição de casos de dengue em Cascavel - PR

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação, do Centro de Ciências Exatas e Tecnológicas da Universidade Estadual do Oeste do Paraná - Campus de Cascavel.

Orientador(a): André Luiz Brun
Coorientador(a): Claudia Brandelero Rizzi

Cascavel–PR

2021

ANGELO JOSÉ ORSSATTO

APLICAÇÃO DE MODELOS ESTATÍSTICOS E DE APRENDIZAGEM DE MÁQUINA NA PREDIÇÃO DE CASOS DE DENGUE EM CASCAVEL - PR

Monografia apresentada como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação, pela Universidade Estadual do Oeste do Paraná, Campus de Cascavel, aprovada pela Comissão formada pelos professores:

Prof. André Luiz Brun (Orientador)
Colegiado de Ciência da Computação, UNIOESTE

Profª. Claudia Brandelero Rizzi (Coorientadora)
Colegiado de Ciência da Computação, UNIOESTE

Profª. Adriana Postal
Colegiado de Ciência da Computação, UNIOESTE

Cascavel, 28 de julho de 2022

Quero para mim o espírito desta frase, transformada a forma para a casar com o que eu sou

Agradecimentos

Agradeço ao SIMEPAR e a Secretaria Municipal de Saúde pelo fornecimento dos dados utilizados nesta pesquisa.

Agradeço à minha família pelo apoio para chegar até aqui, minha mãe Lourdes e meu pai Neuso, que deram tudo de si para proporcionar o melhor para seus filhos. Agradeço aos meus irmãos Fabio e Rafael pelo companheirismo e pelas orientações ao longo dos anos.

Grato também aos amigos, Augusto, Girardi, Fermino, Igor, Emanoel e Valquíria, que fiz durante a graduação, que estão comigo desde o começo dessa jornada, e ao meu parceiro de todas as horas, Luis Norbiato, que continuou comigo mesmo em caminhos distintos porém nunca distantes.

Agradeço com muito amor à minha namorada, Leticia Zago, por todo o carinho, compreensão e atenção nessa reta final, e pela revisão deste texto.

Um agradecimento especial aos meus professores, em particular ao Prof André e a Profª Claudia, que prestaram grande suporte e auxílio ao longo dos anos de trabalho científico.

Aos meus colegas de trabalho do NTI, pela paciência e pelos conhecimentos e experiências compartilhados.

À minha psicóloga, Kettlyn, pelo acompanhamento e tratamento ao longo deste ano.

Agradeço, ainda, à Bateria Primata por ter proporcionado momentos inesquecíveis e por fazer parte da minha história na Universidade.

Quanto vale a vida de qualquer um de nós
Quanto vale a vida em qualquer situação
Quanto valia a vida perdida sem razão
(...)
Quanto vale a vida perto do fim do mês
Quanto vale a vida longe de quem nos faz viver
(...)
São segredos que a gente não conta
E faz de conta que não quer nem saber
São segredos que a gente não conta
Contas que a gente não faz
Coisas que o dinheiro não compra
Perguntas que a gente não faz
Quanto vale a vida?
(Humberto Gessinger)

Resumo

A dengue, que é transmitida pelo mosquito *Aedes aegypti*, se tornou uma grande preocupação urbana, principalmente na ocorrência de surtos da doença. Dessa forma, a previsão do número de casos da dengue em um determinado local adquire uma função social primordial possibilitando a tomada de ações preventivas para um controle efetivo da doença, tanto no combate ao vetor quanto no tratamento dos infectados. O propósito deste trabalho foi avaliar a aplicabilidade de técnicas estatísticas e de aprendizagem de máquina na construção de modelos de predição para o número de casos de dengue no município de Cascavel-PR. Para realização das previsões de casos de dengue, foram avaliadas nove estratégias capazes de realizar predições em séries temporais, sendo três estatísticas e seis modelos de aprendizagem de máquina, dos quais três são abordagens monolíticas e outras três baseadas em *pools* de regressores. A comparação entre os modelos baseou-se nas métricas dos erros médios absolutos (MAE), erros médios quadrados (MSE) e raiz dos erros médios quadráticos (RMSE). O conjunto de dados utilizado nos experimentos compreende a série temporal do número de casos confirmados de dengue referentes ao período de janeiro de 2007 até dezembro de 2020. Além disso, foram empregadas as informações de umidade relativa (%), índice de precipitação (mm) e temperatura média (°C) para o mesmo período. Os resultados dos experimentos realizados mostraram que foi possível realizar a previsão de casos de dengue com os modelos construídos com boa precisão. As abordagens que apresentaram os melhores desempenho foram o SVM (RMSE de 0,094) seguido das duas variações do XGBboost (baseado em árvores e linear) cujos RMSE foram de 2,596 e 4,459, respectivamente. Os piores desempenhos foram observados para os modelos de suavização exponencial (RMSE de 748415) e o *bagging* combinado com o SVM (RMSE de 746662). Acredita-se que os modelos de aprendizagem de máquina, mais especificamente o SVM e o XGBoost, obtiveram melhor performance por combinar o número de casos positivos com as variáveis climáticas. Além disso, observou-se que as estratégias construídas conseguiram estimar com precisão o surto de dengue ocorrido no ano de 2020, cujos valores são bastante discrepantes em relação ao restante da série.

Palavras-chave: Modelos de predição, Séries temporais, XGBoost, SVM, Dengue, Cascavel-PR

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AD	Árvore de Decisão
AM	Aprendizagem de máquina
ARIMA	*Autoregressive Integrated Moving Average*
GTA	Grupo Técnico Consultivo de Imunizações
GRU	*Gated Recurrent Unit*
LSTM	*Long Short-Term Memory*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
MM	Média Móvel
MSE	*Mean Squared Error*
OPAS	Organização Pan-Americana da Saúde
PAHO	*Pan American Health Organization*
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
RNN	*Recurrent Neural Network*
SMC	Sistemas de Múltiplos Classificadores
SE	Suavização Exponencial
SVM	*Support Vector Machine*
XGBoost	*Extreme Gradient Boosting*

Lista de símbolos

$Z(t_n)$	Observações em uma série temporal
t_n	Instante de tempo
h	Horizonte de predição
$\hat{y}$	Valor predito pelo modelo
y_t	Valor de uma observação em um instante de tempo
ϕ_i	Termos descritores da função autorregressiva dependentes da ordem p
θ_i	Termos descritores da função de média móvel dependentes da ordem q
ε	Erro
p	Ordem dos elementos na etapa autorregressiva
d	Grau de diferenciação
q	Ordem da média móvel
A_i	Elementos que compõem a média
n	Número de amostras consideradas
$\hat{y}_{T+1\|T}$	Esperança de um valor predito
α	Parâmetro de suavização
R^m	Plano de dimensão dos valores de entrada
R^o	Plano de dimensão dos valores de saída
X	Conjunto de entrada
x_i	Elementos do conjunto de entrada
W_1	Vetor de pesos na camada de entrada
W_2	Vetor de pesos na camada escondida
b_1	Tendência adicionada à camada escondida
b_2	Tendência adicionada à camada de saída
$\frac{\alpha}{2}\|\|W\|\|_2^2$	Termo de regularização
ζ	Limites de tolerância

Sumário

1 Introdução

A dengue é uma doença infecciosa causada por um vírus da família *Flaviridae*, transmitida principalmente pela picada do mosquito *Aedes aegypti*. Os diferentes sorotipos da patologia (DEN-1 à DEN-5) podem causar enfermidades graves e mortais. Os principais sintomas são febre alta, dores de cabeça, dores no corpo e nas articulações, fraqueza, dor atrás dos olhos e prurido (WHO, 2022). A assistência em saúde visa apenas atenuar a ação dos sintomas. Visto que não existe uma vacina efetiva, a estratégia mais adequada para mitigar os casos de dengue é o combate ao vetor do vírus (RIZZI et al., 2017).

Países tropicais são os mais afetados por esta enfermidade, devido às suas características climáticas, ambientais e socioeconômicas (RIBEIRO et al., 2006) que favorecem a proliferação do *Aedes*, bem como a transmissão do vírus entre a população. No Brasil, foram registradas 502 mortes pela dengue no ano de 2022 e um total de 1.476.486 casos da doença (PAHO, 2022a), até então, com incidência (casos prováveis) de 685,6 por 100.000 habitantes (PAHO, 2022b) e apresentando uma taxa de mortalidade de 0,034% (PAHO, 2022c).

O estado do Paraná totaliza 42.273 diagnósticos positivos de dengue, do início do calendário epidemiológico até dia 29 de março de 2022, sendo que cerca de 80% das notificações são autóctones, ou seja, o paciente contrai a doença no município onde mora (CBNLONDRINA, 2022). A cidade de Cascavel por sua vez, já registrou 7145 casos positivos de dengue no atual ano epidemiológico (período que iniciou em 01/08/2021 e se encerra no dia 31/07/2022) (CGN, 2022). Dado volume de casos foi decretado quadro de epidemia da doença desde 13 de abril de 2022 (G1, 2022a).

A doença ainda é uma preocupação nas diferentes esferas governamentais, tanto federal, estadual ou municipal. Cabe destacar que há um gasto para o sistema público de saúde, proporcional à incidência da doença, variando de US$ 413 a US$ 966, para cada caso hospitalizado (LASERNA et al., 2018).

Segundo Phung et al. (2015), modelos de previsão são ferramentas válidas na preparação

e controle do surto da doença. Com este auxílio, as instituições de saúde poderiam organizar a disponibilidade de recursos materiais e humanos, de forma a minimizar a taxa de mortalidade e otimizar o atendimento aos enfermos.

Um alerta precoce de surtos de dengue pode aumentar a eficácia de campanhas de controle ao vetor e orientar ações preventivas. Assim, intervenções prévias podem retardar e mitigar a intensidade da epidemia e, consequentemente, amenizar seu impacto sanitário e social, além de reduzir a mortalidade, por meio da resposta prévia adequada do sistema público de saúde (GHARBI et al., 2011).

Ao possibilitar a tomada de ações de forma preventiva de controle da doença, a previsão de incidência de dengue adquire um papel social indispensável. Várias abordagens podem ser empregadas para tal, utilizando séries temporais e dados climáticos, focados em uma determinada região geográfica, aplicados em estratégias de aprendizagem de máquina e modelos matemáticos/estatísticos, como os trabalhos de Pham et al. (2016), Braga et al. (2017), Mittelmann e Soares (2017b) e Baquero, Santana e Neto (2018), nos quais foram utilizados o modelo ARIMA, Árvores de decisão, Redes Neurais Artificiais e Rede Neural Recorrente, respectivamente, para realizar as predições de incidência e casos de dengue, frequentes na literatura sobre o tema.

1.1 Objetivos

Com este estudo, objetivou-se implementar e analisar diferentes modelos para estimação do número de casos de dengue em Cascavel - PR, utilizando três abordagens estatísticas e seis estratégias de aprendizagem de máquina, sendo estes: Média móvel; Suavização Exponêncial; ARIMA; *Support Vector Machine*; *Multilayer Perceptron*; *Recurrent Neural Network* com a adoção de células *Long Short-Term Memory* e *Gated Recorrent Unit*; *Random Forest*; *Bagging* sobre modelos de redes neurais, vetores de suporte e árvores de decisão, e *Extreme Boosting* sobre estratégias lineares ou baseadas em árvores de decisão.

Para atingir tal objetivo, mostrou-se necessário realizar a consolidação dos dados a serem estudados, removendo ruídos e tratando informações pertinentes de forma coerente. Realizar o estudo do comportamento dos dados, referente à sazonalidade dos casos de dengue e a correlação com as variáveis climáticas, com diferentes granularidades e deslocamentos das variáveis ao longo da série, com base no ciclo de vida do mosquito.

1.2 Organização do Texto

Além deste texto introdutório, o presente trabalho está subdividido em mais quatro capítulos. O Capítulo 2 descreve os principais conceitos utilizados nesta monografia, incluindo noções básicas sobre séries temporais e predição, os modelos matemáticos e estatísticos e de aprendizagem de máquina listados, cada um descrito separadamente em uma subseção e, por

fim, noções sobre a dengue, como é transmitida, principais sintomas e principais meios para combater o avanço da doença.

O Capítulo 3 descreve a metodologia utilizada para realização do projeto, descrevendo as principais etapas realizadas na obtenção e no tratamento dos dados, as métricas escolhidas e o ambiente onde foram implementados e executados tais modelos. O Capítulo 4 discorre sobre os resultados obtidos com os experimentos, exibindo-os textualmente e graficamente. Além disso, envolve a comparação e avaliação entre os modelos implementados conforme as métricas de erros escolhidas. Por fim, o Capítulo 5 apresenta as conclusões que este estudo permitiu constatar, além de sugestões para continuidade futura para este trabalho.

2

Referencial Bibliográfico

Neste capítulo são abordados os principais tópicos e conceitos explorados nesta pesquisa. Tais conceitos envolvem a classificação dos dados como uma série temporal, os modelos de predição utilizados e a dengue em si, seus impactos no Brasil e na região de Cascavel - PR, juntamente com trabalhos correlatos em que a computação inteligente foi usada como ferramenta de combate à doença.

2.1 Séries Temporais

Uma série temporal é uma coleção de observações feita sequencialmente através do tempo. Pode-se citar exemplos que ocorrem nos mais diversos segmentos, como na economia e na meteorologia. Os métodos de análise temporal constituem uma importante área na estatística (CHATFIELD, 2003).

Os dados de séries temporais e sua respectiva análise assumem uma importância cada vez maior, devido à produção mais volumosa de dados (NIELSEN, 2021). Tal fenômeno, por vezes denominado *Big Data*, é alavancado por meio da digitalização de sistemas, Internet das Coisas (IoT), surgimento de cidades inteligentes, redes sociais, entre outras aplicações presentes mais constantemente na sociedade.

A análise de séries temporais, geralmente se resume à uma questão da causalidade, de como o passado pode influenciar o futuro. As respostas dessa pergunta são tratadas estritamente dentro do campo de atuação, onde está se analisando os dados temporais, não fazem parte do campo geral da análise de série temporal (NIELSEN, 2021). Como resultado disso, várias especialidades contribuíram com inovações ao modo de pensar, sobre os conjuntos temporais, a exemplo da medicina, meteorologia e economia.

2.1.1 Considerações Gerais

Uma série temporal é qualquer conjunto de observações ordenadas no tempo (MORETTIN; TOLOI, 2006). Como por exemplo, valores mensais de pluviosidade, temperatura média diária, ou mesmo o número de casos de dengue por dia em uma cidade. Esses exemplos são categorizados como uma série temporal discreta, obtida através da amostragem contínua em intervalos de tempos iguais Δt.

Obtida uma série temporal $Z(t_1), ..., Z(t_n)$, observada nos instantes $t_1, ..., t_n$, pode-se investigar o mecanismo gerador da série temporal, realizar previsões de valores futuros da série, descrever o comportamento da série, e/ou procurar periodicidades relevantes nos dados (BOX; JENKINS; REINSEL, 2016).

Os métodos tradicionais de análise de séries temporais se preocupam principalmente com a decomposição da série em componentes que representam tendência, sazonalidade, estacionariedade e outras mudanças cíclicas. Qualquer variação restante é interpretada como flutuações irregulares (CHATFIELD, 2003).

A **sazonalidade** está presente em muitas séries temporais, isso se deve ao fato de apresentar um acréscimo do determinado acontecimento, em um período específico do ano (CHATFIELD, 2003). Exemplificando, a medida da temperatura aumenta em época de verão, número de vendas aumenta no verão, desemprego tem maior ocorrência em períodos de inverno.

A **tendência** pode ser definida como uma mudança de longo prazo, no nível médio. Esse tipo de variação ocorre quando uma série exibe um crescimento ou declínio constante, ao longo de vários períodos sucessivos (CHATFIELD, 2000). Já a **estacionariedade**, se dá quando a série desenvolve uma média constante ao redor de um tempo aleatório, implicando que a série se mantém estável, refletindo um equilíbrio (MORETTIN; TOLOI, 2006). Séries podem ser estacionárias durante um curto período de tempo dentro da série.

2.1.2 Predição

Prever os valores futuros de uma série temporal é um problema importante em diversas áreas. As previsões podem ser feitas de forma subjetiva usando julgamento, intuição, conhecimento comercial e qualquer outra informação relevante (CHATFIELD, 2003).

Em geral, a predição é feita sobre um período de observação onde calcula-se valores $Z(t_{n+h})$ que são desconhecidos na série, na qual, h refere-se ao horizonte de previsão (MILLS, 2019). Previsões univariadas de uma variável são baseadas em modelos dependentes de seus valores no presente e passado na série temporal, então Z_{n+h} depende exclusivamente dos valores da série $Z_n, Z_{n-1}, ..., Z_1$. Esses métodos são tipicamente chamados de métodos de projeção. Previsões multivariáveis envolvem outras séries adicionais no estudo de uma variável, chamadas de variáveis preditoras ou explicativas.

Uma grande variedade de procedimentos para realizar predições em séries temporais estão disponíveis na literatura, e é importante ressaltar que nenhum método é universalmente aplicável, mas sim, é preciso escolher o mais apropriado, dadas as devidas condições (CHATFIELD, 2000). A escolha da estratégia de predição inclui questões relacionadas a como ela será utilizada, o tipo da série temporal e suas propriedades, e quantas observações passadas estão disponíveis. Tais tópicos serão abordados dentro de cada subseção correspondente a cada modelo descrito neste trabalho.

A análise deve fazer perguntas para obter informações suficientes, deve esclarecer os objetivos na produção das previsões e deve descobrir como a previsão será usada, tudo isso aliado ao contexto em que a análise está sendo realizada (CHATFIELD, 2000).

Três modelos estatísticos para predição em séries temporais são discutidos na seção 2.2 onde serão abordadas as principais estratégias que cada um utiliza. Na seção 2.3, será discutido a utilização de modelos de aprendizagem de máquina no escopo da análise e predição de séries temporais.

2.2 Modelos Estatísticos

Os modelos utilizados para descrever séries temporais são processos controlados por leis probabilísticas, isto é, processos estocásticos. A construção de modelos depende de vários fatores, tais como o comportamento do fenômeno estudado ou o conhecimento prévio de sua natureza (MORETTIN; TOLOI, 2006). Na prática, depende também da disponibilidade de *softwares* adequados para tal estimação.

Nas sessões seguintes, serão introduzidos os modelos estatísticos utilizados neste trabalho, bem como suas particularidades durante a modelagem.

2.2.1 Média Móvel

A média móvel fornece um método simples para suavizar o passado histórico dos dados (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997). Essa métrica considera apenas a variável objetivo para realizar sua predição, ou seja, apenas utiliza os valores descritos pela série temporal. Médias móveis são geralmente calculadas para identificar a direção da tendência de um evento (FERNANDO, 2021).

O termo é usado para descrever o procedimento em que cada média é calculada eliminando a observação mais antiga e incluindo a próxima observação. A média se move através da série temporal, até que o ciclo de tendência seja calculado em cada observação para todos os elementos disponíveis (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997).

A média móvel simples é calculada pela média aritmética de um determinado conjunto de valores, durante um período de tempo especificado:

$$MM = \frac{y_1 + y_2 + \cdots + y_n}{n} \tag{1}$$

em que y_i são os elementos da série temporal e n corresponde ao período de tempo anterior à estimação. O valor de n incluída em uma média móvel afeta a suavidade da estimativa resultante (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997). Portanto, quanto maior o valor de n mais suave será a curva de resultados.

A média móvel é amplamente usada na análise técnica, tanto para detectar uma mudança ou para confirmar suspeitas de uma mudança que possa estar em andamento (MORETTIN; TOLOI, 2006).

2.2.2 ARIMA

O modelo ARIMA (*Autoregressive Integrated Moving Average*) provê uma abordagem para previsão de séries temporais que integra o modelo autorregressivo e a média móvel. O modelo de suavização exponencial (descrito na subseção 2.2.3) e ARIMA são as duas abordagens mais utilizadas para a previsão de séries temporais e fornecem abordagens complementares para o problema. Enquanto que os modelos de suavização exponencial são baseados na descrição da tendência e sazonalidade dos dados, os modelos ARIMA visam descrever as autocorrelações nos dados (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997).

Antes de descrever o modelo ARIMA em si, é importante detalhar brevemente os modelos autorregressivos e os de média móvel, que são incorporados no modelo ARIMA.

Em um modelo de autorregressão, a variável de interesse é predita utilizando uma combinação linear de valores anteriores desta variável. Portanto, um modelo autorregressivo $AR(p)$, é determinado pela Equação 2.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \tag{2}$$

onde $y_{t-1}, \ldots, y_{t-p}$ são as observações passadas da série, ϕ_i é o coeficiente que deve-se determinar para estabelecer uma relação entre y_t e y_{t-1} e ε_t corresponde ao erro

Os modelos autorregressivos são notavelmente flexíveis ao lidar com uma ampla gama de diferentes padrões de séries temporais (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997).

Um modelo de média móvel que utiliza erros de predições passadas para obter valores atuais $MA(q)$ é descrito na Equação 3 a seguir. Nela, ε_t corresponde ao erro de cada predição anterior.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

Neste modelo, cada valor de y_t pode ser considerado como uma média móvel ponderada dos erros de predições passadas.

Combinando, então, o modelo autorregressivo (Equação 2) e o modelo de média móvel (Equação 3), obtém-se o modelo ARIMA completo, descrito na Equação 4.

$$y_t = c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{4}$$

Os "preditores" do lado direito incluem ambos os valores defasados de y_t e erros retardados. O modelo também é descrito como ARIMA(p,d,q), onde p é a ordem dos elementos provenientes da etapa autorregressiva, d é o grau de primeira diferenciação envolvido e q a ordem da regressão por média móvel dos erros (HYNDMAN; ATHANASOPOULOS, 2018).

2.2.3 Suavização Exponencial

A suavização exponencial produz previsões, nas quais as médias ponderadas de observações passadas recebem pesos que decaem exponencialmente à medida que as observações envelhecem. Em outras palavras, quanto mais recente a observação, maior o peso associado a ela (HYNDMAN; ATHANASOPOULOS, 2018).

O valor do elemento previsto $\hat{y}_{T+1|T}$ é dado pela Equação 5.

$$\hat{y}_{t+1|t} = \alpha y_t + \alpha(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots \alpha(1-\alpha)^t y_1 \tag{5}$$

onde $0 \leq \alpha \leq 1$ é o parâmetro de suavização, que controla a taxa na qual os pesos diminuem relacionados as observações $y_1, \ldots y_T$.

Quando um pequeno valor de α é escolhido, a previsão inicial desempenha um papel mais proeminente do que quando um α maior é usado. Se o parâmetro de suavização não estiver próximo de zero, a influência do processo de inicialização torna-se rapidamente menos significante com o passar do tempo. No entanto, se α for próximo de zero, o processo de inicialização pode desempenhar um papel significativo por um período longo de tempo (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997).

Essa abordagem gera previsões confiáveis de forma rápida e para uma ampla gama de séries temporais, possuindo ou não sazonalidade e tendência, o que se torna uma grande vantagem, aumentando sua aplicabilidade (HYNDMAN; ATHANASOPOULOS, 2018).

2.3 Modelos de Aprendizagem de Máquina

Métodos de aprendizado de máquina podem ser eficazes em problemas de previsão de séries temporais mais complexos, com múltiplas variáveis de entrada, relacionamentos não

lineares complexos e dados ausentes (BROWNLEE, 2018). Dado este fato, neste trabalho foram elencados seis modelos de aprendizagem de máquina, sendo três deles modelos monolíticos, ou seja, realizam as predições por sí próprios, e os outros três, consistem em modelos de agrupamento, no qual utilizam de vários classificadores ou regressores para definir uma opinião final (TAN; STEINBACH; KUMAR, 2009).

Os modelos e estratégias utilizados serão descritos nas seguintes seções cujas escolhas foram baseadas principalmente nos trabalhos de Pham et al. (2016), Braga et al. (2017), Guo et al. (2017) e Muhilthini et al. (2018).

2.3.1 *Multilayer Perceptron*

O Perceptron de Múltiplas Camadas (*Multilayer Perceptron* - MLP) é um algoritmo de aprendizado supervisionado que aprende uma função $f : R^m \rightarrow R^o$ por treinamento de conjunto de dados, onde m é o número de dimensões para entrada (m refere-se ao número de atributos do conjunto) e o corresponde ao número de dimensões para saída (Scikit-learn developers, 2021c). Dado um conjunto $X = x_1, x_2, ..., x_n$ e um alvo y, o modelo visa aprender um aproximador de função não linear para a tarefa de classificação (neste caso o corresponde ao número de classes possíveis) ou regressão (TAN; STEINBACH; KUMAR, 2009). Neste caso o algoritmo retorna um valor contínuo e não apenas um entre vários possíveis como na classificação

A principal vantagem do MLP é a capacidade de aprender modelos não lineares, contudo apresenta uma função de perda não convexa nas suas camadas escondidas e é sensível ao nível de variação dos atributos (TAN; STEINBACH; KUMAR, 2009). Uma boa prática para diminuir essa sensibilidade é fazer a padronização dos dados, ou seja, para cada valor de entrada x transformá-lo em um valor entre $[0, 1]$, onde o maior x é igual a 1, em consequência, o menor x torna-se 0.

O MLP, quando aplicado a problemas de regressão, é treinado usando retropropagação utilizando a função identidade como a de ativação. A função de perda utilizada é o erro quadrado e a saída é um conjunto de valores contínuos. Além disso, também usa um parâmetro α para regularização que ajuda a evitar o *overfitting* na função de perda (Scikit-learn developers, 2021c).

Dado um conjunto de treinamento $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, um neurônio em uma camada oculta aprende a Equação 6.

$$f(x) = W_2 g(W_1^T x + b_1) + b_2 \tag{6}$$

onde W_1 e W_2 são vetores que representam os pesos na camada de entrada e nas camadas escondidas, respectivamente, e b_1, b_2 correspondem às tendências adicionadas à camada escondida e à camada de saída, respectivamente (TAN; STEINBACH; KUMAR, 2009). Na regressão, a função de saída corresponde a $f(x)$ e a função de ativação é a função identidade (Scikit-learn developers, 2021c).

Para a função de perda em problemas de regressão, o MLP utiliza a função do erro quadrático, conforme descrito na Equação 7:

$$Loss(\hat{y}, y, W) = \frac{1}{2}||\hat{y} - y||_2^2 + \frac{\alpha}{2}||W||_2^2 \tag{7}$$

sendo $\frac{\alpha}{2}||W||_2^2$ um termo de regularização que penaliza modelos complexos com α controlando a magnitude dessa penalização. Começando com pesos de valor aleatório, o MLP minimiza a função de perda atualizando continuamente esses pesos (TAN; STEINBACH; KUMAR, 2009). Após computar o valor de perda, realiza-se a propagação em sentido contrário, da camada de saída em direção às camadas anteriores, fornecendo um peso de valor atualizado destinado a diminuir a perda (Scikit-learn developers, 2021c).

O algoritmo finaliza sua execução quando atinge um número máximo de iterações ou quando a melhora no valor da perda está abaixo de um certo patamar especificado a priori.

2.3.2 *Support Vector Machine*

As Máquinas de Vetores de Suporte (*Suport Vector Machines* - SVMs) são métodos de aprendizado supervisionado, usados tanto para problemas de classificação ou regressão, como neste caso, construindo um hiperplano ou um conjunto de hiperplanos em um espaço dimensional mais elevado (BISHOP, 2006)

Na regressão, o objetivo é encontrar uma função $f(x)$, que tenha no máximo ζ desvios dos alvos reais y_i para todo o conjunto de treinamento. Em outras palavras, erros não são considerados desde que sejam menores que ε, mas não aceita-se qualquer desvio maior do que este (SMOLA; SCHOLKOPF, 2004).

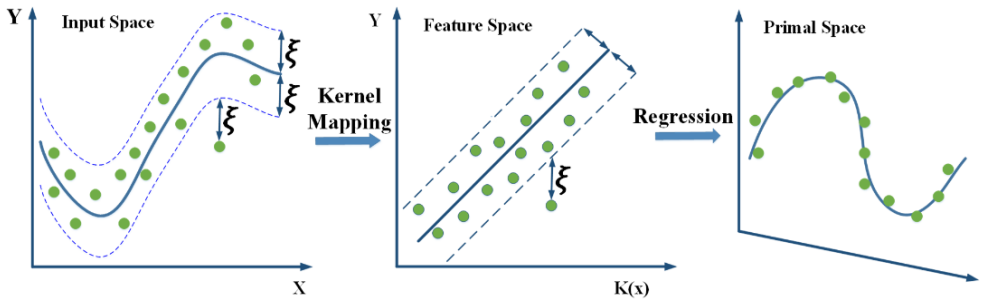
A exemplo, na Figura 1, encontra-se a função $f(x)$ que melhor representa o comportamento das amostras de treinamento, com limite inferior $f(x) - \zeta$ e limite superior $f(x) + \zeta$.

Após a definição das funções, realiza-se o mapeamento do *kernel* (núcleo) para transformação do conjunto no hiperplano superior, caso necessário. As funções mais comuns para o *kernel* do SVM são funções lineares, polinomiais, de bases radiais (rbf) ou sigmóides (Scikit-learn developers, 2021e).

Por fim, os valores são preditos pelo SVM com a regressão, encaixando-se à curva descrita na execução de sua primeira etapa.

Em geral, os SVMs são eficazes em espaços de alta dimensão ou quando o número de dimensões é maior que o número de amostras. Além disso, é eficiente em termos de memória, por utilizar o subconjunto de pontos de treinamento na função de decisão, também chamados de vetores de suporte (Scikit-learn developers, 2021e). A versatilidade deste modelo chama atenção, devido à possibilidade de utilizar diversas funções de *kernel* (BISHOP, 2006).

Figura 1 – Processo de regressão do SVM



Fonte: (MORADZADEH et al., 2020)

Contudo, os SVMs não oferecem estimativas de probabilidade de forma direta, elas são calculadas utilizando validação cruzada. A função de *kernel* também pode ser sensível ao número de recursos que utiliza em relação ao número de amostras (Scikit-learn developers, 2021e).

2.3.3 Árvores de Decisão

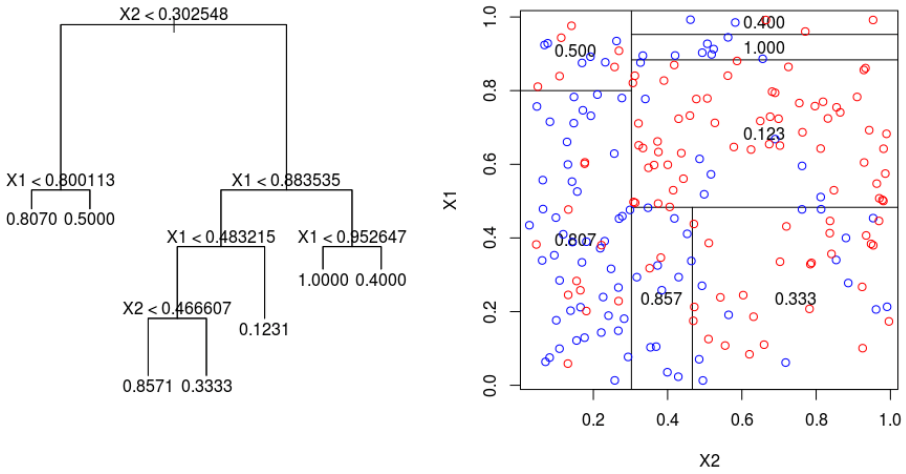
Árvores de decisão (ADs) são um método de aprendizado supervisionado não paramétrico, cujo objetivo é criar um modelo que preveja o valor de uma variável por meio de regras inferidas, com base em recursos advindos dos próprios dados (Scikit-learn developers, 2021a).

Dado valores de treinamento $x_1, x_2, \ldots, x_l$ e um valor objetivo y, uma árvore de decisão particiona recursivamente o espaço de recursos de forma que as amostras com os mesmos rótulos ou valores de destino semelhantes sejam agrupadas (BREIMAN et al., 1984). Este processo está ilustrado na Figura 2, no qual a árvore é particionada até que atinja o número de folhas especificado ou que atinja um particionamento satisfatório, a partir de regras criadas segundo o próprio conjunto de treinamento, nas quais são demonstradas do lado esquerdo da figura e o lado direito exemplifica a determinação dos valores dentro da subdivisão do espaço onde estão dispostos os elementos.

Com a árvore definida, o modelo irá realizar a verificação de qual quadro cada valor objetivo encaixa-se a partir das regras moldadas.

Pode-se destacar que as árvores de decisão apresentam um modelo simples de interpretar, que não requer muita preparação dos dados, e é possível validar o modelo utilizando testes estatísticos. Entretanto, árvores podem ser instáveis, pois uma pequena variação no conjunto de dados pode resultar em uma árvore completamente distinta (BREIMAN et al., 1984). As árvores podem criar modelos muito específicos que não generalizam bem os dados, e este é o principal motivo para a utilização de um conjunto de árvores, ou seja, a floresta aleatória (Scikit-learn developers, 2021a).

Figura 2 – Processo de regressão empregando-se uma Árvore de Decisão



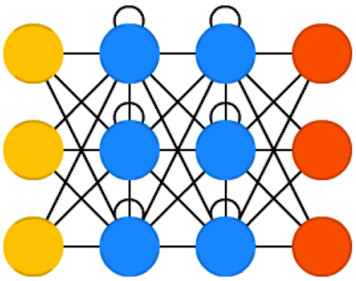
Fonte: (LE, 2018)

2.3.4 *Recurrent Neural Network*

As Redes Neurais Recorrentes (*Recurrent Neural Network* - RNN) são uma classe de redes neurais poderosas para modelar dados em sequência, como séries temporais (KERAS, 2022). Esse tipo de rede neural possui conexões entre suas passagens e através do tempo. Os neurônios são alimentados com informações da camada anterior a ele e informações de si mesmo da passagem anterior. Com isso, a alimentação de uma RNN depende da ordem de suas entradas (VEEN, 2016)

A Figura 3, representa a estrutura básica de uma rede neural recorrente, na qual os círculos amarelos representam as entradas, os círculos azuis representam as células recorrentes nas quais existe uma ligação para si próprio representando a auto-alimentação de dados, e os círculos laranjas representam a saída.

Figura 3 – Estrutura de uma Rede Neural Recorrente



Fonte: (VEEN, 2016)

Os principais pontos negativos das RNNs envolvem a lentidão em que os cálculos podem

ser realizados, a não consideração de entradas futuras para tomada de decisões e o problema de gradiente de fuga, onde os gradientes usados para calcular a atualização dos pesos pode ficar muito próximo de zero, impedindo que a rede aprenda novos pesos (GOODFELLOW; YOSHUA; COURVILLE, 2016). Para resolver este último problema, foram introduzidas as camadas *Long Short-Term Memory* (LSTM) à rede neural recorrente, discutida na subseção 2.3.4.1.

2.3.4.1 *Long Short-Term Memory*

A camada *Long Short-Term Memory* (LSTM) é utilizada em arquiteturas de redes recorrentes em conjunto com um algoritmo de aprendizado baseado em gradiente apropriado. A LSTM foi projetada para superar erros de fuga de gradiente, podendo aprender a preencher intervalos de tempo em até 1000 passos, mesmo no caso de sequências de entradas ruidosas e incompressíveis, sem perda em curto espaço de tempo (HOCHREITER; SCHMIDHUBER, 1997).

Cada neurônio possui uma célula de memória e três conexões: entrada, saída e esquecimento. A função dessas conexões é resguardar a informação, interrompendo ou permitindo seu fluxo. A conexão de entrada determina quanto da informação da camada anterior é armazenada na célula. A conexão de saída determina quanto de informação da camada anterior será repassada para a próxima camada. Já a conexão do esquecimento, considera um possível descarte de uma informação que pode não ser mais útil ao longo do tempo (VEEN, 2016).

Combinando assim, redes neurais recorrentes com células do tipo LSTM é possível criar modelos mais robustos e que permitem uma maior generalização dos dados, tornando a rede mais aplicável e mais eficiente (BROWNLEE, 2018).

2.3.4.2 *Gated Recurrent Unit*

As *Gated Recorrent Units* (GRU) são variações das células LSTM, descritas na seção anterior. A sua diferença está no número de conexões e a maneira que estão ligadas. A GRU possui uma conexão de atualização, que determina o quanto manter da informação do estado anterior e o quanto que será repassado para o próximo estado (VEEN, 2016).

Com essa arquitetura, é mais fácil para cada unidade lembrar da existência de um fato específico no conjunto de entrada ao longo da série de passos (CHUNG et al., 2014). Além disso, a conexão de atualização cria uma espécie de atalho que permitem os erros serem retropropagados mais facilmente (HOCHREITER; SCHMIDHUBER, 1997).

Apesar das diferenças entre as duas arquiteturas, é difícil concluir qual célula performará de maneira mais eficiente, de maneira geral (BAHDANAU; CHO; BENGIO, 2014). Com isso, recomenda-se criar modelos de rede neural recorrente que utilize tanto células LSTM quanto células GRU (BENGIO; BOULANGER-LEWANDOWSKI; PASCANU, 2013).

As estratégias de regressão apresentadas até então realizam suas predições a partir de

uma única instância do modelo. Tais abordagens são ditas monolíticas. Apesar de apresentarem resultados interessantes, em problemas de maior variabilidade os modelos monolíticos podem não ser capazes de cobrir todas as variações do problema.

Uma alternativa às estratégias monolíticas, é unir o esforço de vários regressores para tentar mitigar o problema da cobertura da variabilidade do espaço de regressão de forma a minimizar os erros e aumentar o desempenho e precisão desses sistemas. Essa é a ideia dos Sistemas de Múltiplos Classificadores (SMC), também conhecidos como *Ensemble Learning* (KITTLER, 1998; GUNES et al., 2003). Nesta pesquisa são adotados os sistemas compostos de múltiplos regressores.

Nas próximas seções são apresentadas as técnicas adotadas para melhorar a precisão dos regressores agregando previsões de múltiplas instâncias de determinado modelo. Tal processo consiste em criar um conjunto de regressores básicos a partir dos dados de treinamento e executar a predição recebendo um voto (classificação) ou média (regressão) sobre as opiniões dadas por cada modelo (TAN; STEINBACH; KUMAR, 2009).

Os métodos de grupo são atraentes principalmente porque são capazes de impulsionar modelos “fracos“ e diminuem significativamente a taxa de erro dos modelos em geral (ZHOU, 2012). Duas estratégias bastante difundidas, o *Bagging* e o *Boosting*, são criados pela reamostragem dos dados originais, manipulando o conjunto de treinamento com o objetivo de formar diferentes conjuntos de aprendizado. Essa estratégia permite que modelos sejam especialistas em diferentes regiões do espaço do problema. A Floresta aleatória, ao contrário das abordagens anteriores, é um método de grupo que manipula as características do conjunto de treinamento (e não as instâncias) e utiliza árvores de decisão como regressor base (TAN; STEINBACH; KUMAR, 2009). Os modelos citados serão descritos nas próximas seções.

2.3.5 *Random Forest*

Uma floresta aleatória (*Random Forest* - RF) é um meta estimador que ajusta várias árvores de decisão de classificação ou regressão em várias subamostras do conjunto de dados e utiliza um critério definido para melhorar a precisão preditiva e controlar o *overfitting* (Scikit-learn developers, 2021d).

Em florestas aleatórias, cada árvore do conjunto é construída a partir de uma amostra retirada do conjunto de treinamento. Além disso, ao dividir cada nó durante a construção de uma árvore, a melhor divisão é encontrada entre todos os recursos de entrada. O objetivo dessa fonte de aleatoriedade é diminuir a variância do estimador florestal (Scikit-learn developers, 2021b).

A aleatoriedade inerente às florestas produz árvores com erros de previsão um tanto independentes e, ao fazer uma média das previsões, alguns erros podem ser cancelados. Desse modo, as florestas aleatórias atingem uma variação reduzida com a estratégia de combinação de diversas árvores (Scikit-learn developers, 2021b). Na prática, isso resulta em um modelo

globalmente melhor, ao custo de um aumento de processamento.

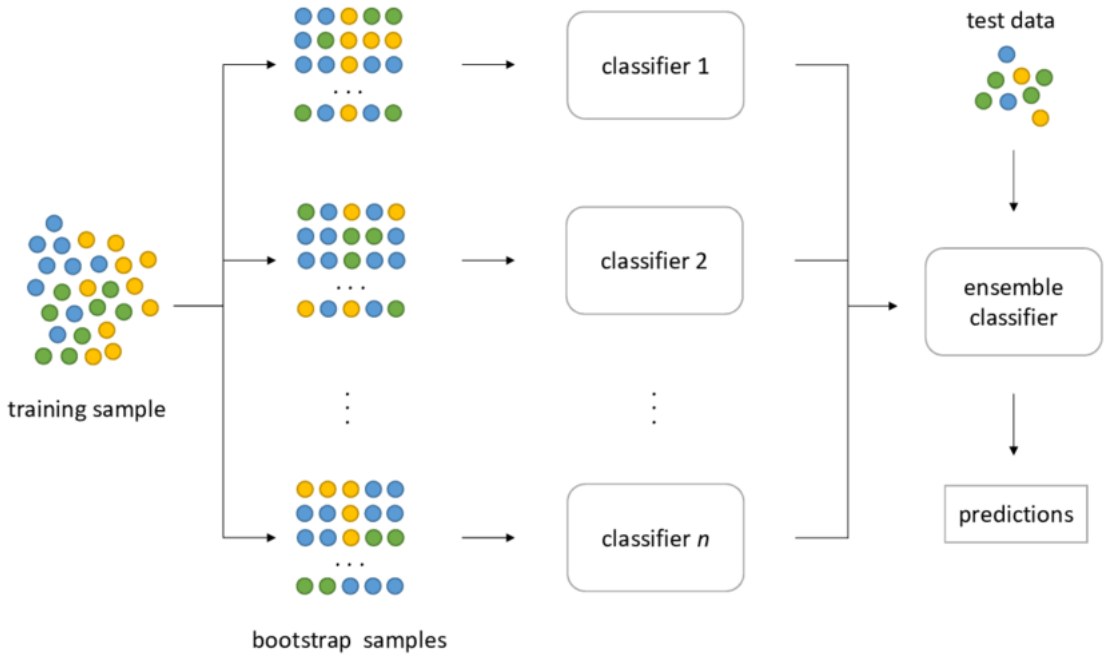
Para compreender melhor esse método, a subseção 2.3.3 explica como é o processo de uma árvore de decisão, mais especificamente voltada para problemas de regressão.

2.3.6 *Bagging*

O *Bagging* (*Bootstrap Aggregatig*), é uma técnica que cria subconjuntos a partir de amostras selecionadas de um conjunto de dados de acordo com uma distribuição uniforme de probabilidade (BREIMAN, 1996). Cada amostra pode possuir o mesmo tamanho dos dados originais ou representar um percentual do conjunto original, visto que utiliza substituições, ou seja, um elemento pode estar presente diversas vezes no conjunto, enquanto outros podem ser omitidos (TAN; STEINBACH; KUMAR, 2009).

Seja n o número de amostras, o *Bagging* cria n subconjuntos e treina um classificador (ou regressor) sobre cada subconjunto formado (TAN; STEINBACH; KUMAR, 2009). Após treinar os n modelos, agrega-se suas previsões individuais, por votação ou por média, a fim de formar uma previsão final (BREIMAN, 1996). Em outras palavras, treina-se n modelos individuais de forma paralela, onde cada modelo é treinado em um subconjunto dos dados determinado de forma randômica, assim como ilustrado na Figura 4.

Figura 4 – Processo de classificação e regressão do *Bagging*



Fonte: (GALDI; TAGLIAFERRI, 2018)

O *Bagging* melhora o erro de generalização reduzindo a variância dos modelos de base e seu desempenho depende da estabilidade do modelo usado. Já que esta técnica não favorece

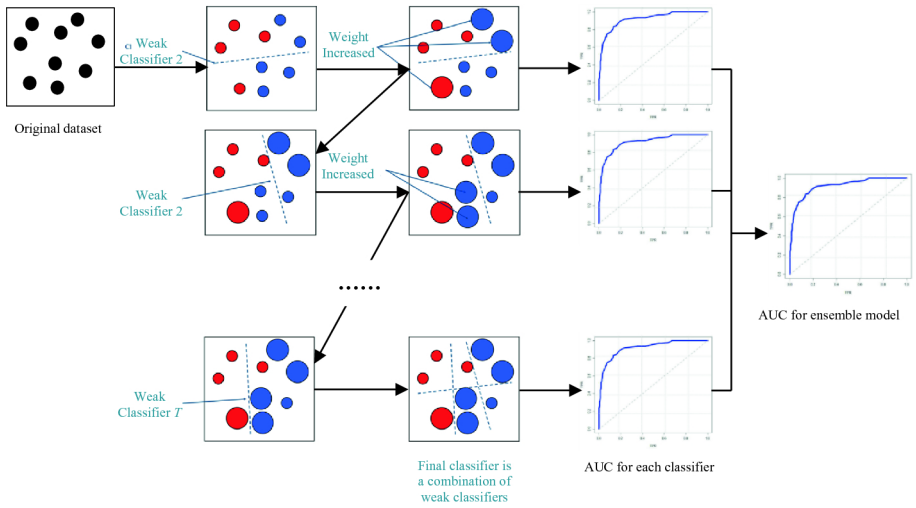
qualquer instância em específico, a estratégia torna-se menos suscetível ao *overfitting* quando aplicada a bases que contenham muitos ruídos (TAN; STEINBACH; KUMAR, 2009).

2.3.7 *Gradient Boosting*

O *Boosting* é um procedimento iterativo usado para alterar, de maneira adaptativa, a distribuição de exemplos de treinamento, de modo que os modelos de base enfonquem em exemplos que sejam mais complexos (TAN; STEINBACH; KUMAR, 2009). O algoritmo atribui um peso para cada exemplo de treinamento, que pode ser usado para construir um conjunto de amostras a partir dos dados originais ou podem ser utilizados como auxílio, a fim de identificar tendência na direção de exemplos de pesos mais altos (TAN; STEINBACH; KUMAR, 2009).

Inicialmente é formado o primeiro subconjunto de forma aleatória equiprovável. As instâncias que formam esse conjunto são submetidas a um classificador (ou regressor) e então têm seu peso ajustado. Assim, como ilustrado na Figura 5, em uma iteração seguinte, as instâncias têm probabilidades distintas de serem escolhidas para formar o novo subconjunto (ZHOU, 2012).

Figura 5 – Processo de *boosting*



Fonte: (Data Science Team, 2020)

A cada iteração os pesos das amostras são ajustados para assim treinar o próximo classificador ou regressor, e são coletados as predições realizadas. Ao final do processo, as predições são analisadas por meio de votação ou por média, a fim de formar uma previsão final, processo similar ao *bagging* (BREIMAN, 1996).

O *Gradient Boosting* constrói um modelo aditivo de forma progressiva que permite a otimização de funções de perda. Em cada estágio, uma árvore de regressão é ajustada no gradiente usando qualquer função de perda diferenciável arbitrária (FRIEDMAN, 2002).

2.4 Dengue

A dengue é uma doença infecciosa causada por um vírus pertencente à família *Flaviviridae*. O vírus da dengue apresenta quatro sorotipos: DENV-1, DENV-2, DENV-3, DENV-4 e DENV-5 (Fiocruz, 2013). Os principais sintomas da doença são febre, dor de cabeça, dores pelo corpo e náuseas. Outros sintomas como, o aparecimento de manchas vermelhas na pele, sangramentos no nariz e nas gengivas, dor abdominal e vômitos podem indicar um alarme para dengue hemorrágica (Biblioteca Virtual em Saúde do Ministério da Saúde, 2007). Esse é um quadro grave, no qual é preciso atenção imediata, por se tratar de um estado que pode ser fatal.

O principal vetor transmissor da dengue é o mosquito *Aedes aegypti*, que possui um comportamento estritamente urbano. Qualquer epidemia de dengue está diretamente relacionada à concentração da densidade do mosquito (Fiocruz, 2013). O principal método para controlar ou prevenir a transmissão do vírus da dengue é combater o mosquito vetor cobrindo, esvaziando e limpando recipientes domésticos que possam armazenar água. Por isso, é importante conhecer os hábitos do mosquito, a fim de combatê-lo como forma de prevenção da doença. O *Aedes* está apto a infectar um ser humano em torno de 20 a 30 dias (WHO, 2022).

Em relação à vacina contra dengue disponível no mercado, Dengvaxia CYD-TDV, sua aplicação em uma estratégia de triagem pré-vacinação ainda exige uma avaliação cuidadosa (WHO, 2022). Por este motivo, ela não faz parte do quadro de vacinas ofertadas à população.

2.4.1 Região de Cascavel - PR e a dengue

A cidade de Cascavel está localizada nas coordenadas 24º57’21"S / 53º27’19"W, no estado do Paraná, região sul do Brasil. Possui uma população estimada em 336.073 habitantes em uma área de 2.103,123km^2 e densidade demográfica de 136,23 hab/km^2, segundo dados disponibilizados pelo IBGE (IBGE, 2021).

A extensão territorial de Cascavel está situada no Terceiro Planalto do Paraná e no encontro de três unidades hidrográficas do Paraná, são elas, a bacia do Rio Iguaçu, do Rio Piquiri e do Rio Paraná. Na Figura 6 é apresentado um mapa do estado do Paraná, com o município de Cascavel em destaque.

Segundo dados disponíveis no portal da Prefeitura Municipal de Cascavel (Prefeitura Municipal de Cascavel, 2021), o Instituto de Terras, Cartografia e Geociências (ITCG) identificou que a cidade enquadra-se nas categorias Cfa (Clima Subtropical úmido), Cfb (Clima Oceânico Temperado) e Cfa/Cfb segundo a classificação de Köppen. Tais categorias indicam que o município não possui uma estação seca e o verão pode ser quente ou fresco. A vegetação nativa que encobre a região de Cascavel é a Mata de Araucárias, contudo grande parte devastada, devido à atividade intensa de agricultura e à expansão urbana (Prefeitura Municipal de Cascavel, 2021).

O índice de infestação de Dengue no município chegou a 5,3% em janeiro de 2020,

Figura 6 – Mapa do Paraná com destaque no município de Cascavel

Fonte: (WIKIPEDIA, 2022)

segundo o Controle de Endemias da Secretaria de Saúde de Cascavel (G1, 2022b). O valor indicado pela Organização Mundial de Saúde é de até 1%. A cidade também apresentou 6.681 casos confirmados entre julho de 2019 e junho de 2020, com boletim epidemiológico que ultrapassava 10 mil notificações (Prefeitura Municipal de Cascavel, 2020).

Frequentemente, mutirões são realizados no município a fim de diminuir a infestação e prevenir novos casos. Geralmente, essas ações são mais centradas em bairros que apresentam maior índice (G1PR, 2021).

2.5 Trabalhos Correlatos

Encontra-se, na literatura, diversas pesquisas, cujos objetivos consistem em estimar o número de casos de dengue utilizando estratégias computacionais, com aprendizagem de máquina e inteligência artificial, estratégias baseadas na estatística, com regressão linear e modelos auto-regressivos. Dentre esses, pode-se destacar os trabalhos de Mittelmann e Soares (2017a), Mittelmann e Soares (2017b) e Baquero, Santana e Neto (2018) que são aplicados no Brasil, respectivamente nas cidades de Guarulhos-SP, Itajaí-SC e São Paulo-SP. Destacam-se também estudos aplicados em outras regiões fora do país, como a pesquisa de Guo et al. (2017) aplicado em Guangdong na China, o de Azhar, Marina e Anwar (2017) com foco em Denpasar, na Indonésia. Podemos citar também os trabalhos de Phung et al. (2015) aplicado em Can Tho no Vietnã e de Laureano-Rosario et al. (2018) realizado em Yucatán (México) e San Juan (Porto Rico).

Um ponto que é comum na maioria dos estudos realizados na área é a utilização de variáveis climáticas juntamente com séries históricas, com o número de casos de dengue já registrados, para auxiliar na predição. Essas variáveis incluem, em grande parte, a precipitação,

umidade relativa e temperatura. Trabalhos como os de Pham et al. (2015), Zhu, Hunter e Jiang (2016), Laureano-Rosario et al. (2018) e Ahmad et al. (2018) entretanto, utilizam outros fatores ambientais e sociais em adição às variáveis climáticas e aos casos de dengue, como o índice de vegetação, temperatura da superfície terrestre e marítima, índice de poluição do ar, tamanho da população e densidade populacional.

Os estudos de Lee, Yang e Lin (2015) e Carlos, Nogueira e Machado (2017) realizaram predições utilizando *data mining*, advindas de diversas bases de dados (*Big data*) e também de redes sociais, como o *Twitter*[1].

A utilização de variáveis climáticas acarreta na restrição geográfica dos estudos, normalmente realizados em uma cidade ou região metropolitana, de países como Brasil, China, Indonésia e Malásia. Tais países são denominados subdesenvolvidos economicamente e ambientalmente classificados como de clima tropical.

Baseado no que foi registrado na literatura, justifica-se a escolha e utilização das variáveis climáticas, aplicadas em uma região específica, deste trabalho que espera-se colaborar com outros futuros.

[1] Disponível em: <https://about.twitter.com/pt>

3
Metodologia

Neste capítulo, é descrito como ocorre o fluxo de execução do projeto. Com uma abordagem metodológica de caráter mais experimental, o trabalho se divide em duas grandes etapas. A primeira consiste na obtenção e tratamento dos dados que foram utilizados, enquanto que a segunda abrange o treinamento dos modelos, sua avaliação e análise dos resultados.

A Figura 7 ilustra o fluxograma da realização do projeto como um todo. Nela são apresentadas as etapas desenvolvidas, desde a obtenção e preparação dos dados, treinamento dos modelos de predição, até a avaliação destes.

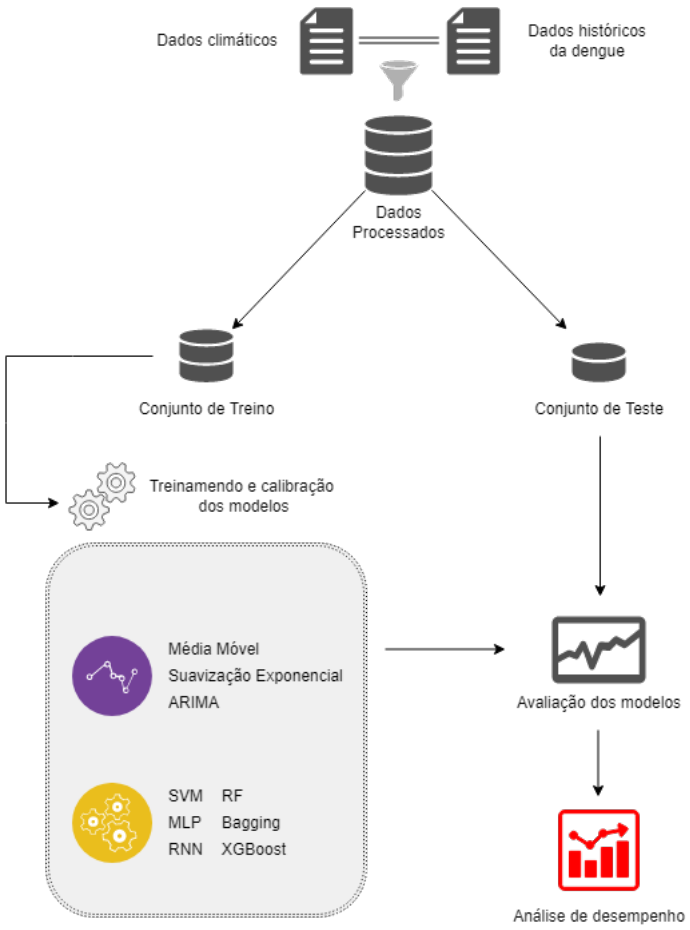
Primeiramente, realizou-se a filtragem dos dados obtidos em relação a dengue e ao clima, incluindo neste processo a remoção e tratamento de possíveis ruídos presentes em ambas as bases. Após a preparação dos dados, o conjunto foi divido em subconjuntos de treino e de teste, sendo o primeiro usado para treinar os modelos: ARIMA, Suavização Exponencial, Média Móvel, *Support Vector Machine*, *Multilayer Perceptron*, Rede Neural Recorrente, Floresta Aleatória, *Bagging* e *Extreme Boosting*. Já o conjunto de teste é utilizado para avaliar o desempenho dos modelos calibrados, para assim concluir sobre a predição de casos de dengue a partir dos resultados apresentados. Maiores detalhes de cada etapa são apresentadas nas seções seguintes.

3.1 Ambiente de execução dos procedimentos e tecnologias utilizadas

Antes de explicitar mais detalhes das etapas do projeto, é importante detalhar as tecnologias utilizadas nesta pesquisa, as quais foram cruciais para o desenvolvimento dos modelos preditivos e a análise dos dados.

Os procedimentos para realização deste trabalho foram implementados utilizando Python

Figura 7 – *Pipeline* da realização da predição dos casos de dengue



Fonte: o autor

como linguagem de programação e executados em ambiente Google Colaboratory[1], também chamado de Colab. O Google Colab é um serviço gratuito de nuvem, desenvolvido para incentivar a pesquisa de Aprendizado de Máquina e Inteligência Artificial. Ao executar experimentos neste ambiente, os códigos ficam livres de contextos dependentes da máquina local e de sistemas operacionais.

Pacotes externos foram utilizados na execução das etapas do projeto. Para o processamento e visualização dos dados bem como para a avaliação das predições, foram utilizadas as bibliotecas Matplotlib[2], Numpy[3], Pandas[4] e Seaborn[5]. Já para o treinamento dos modelos, foram utilizadas

[1] Disponível em: <https://colab.research.google.com/?utm_source=scs-index>
[2] Disponível em: <https://matplotlib.org/stable/>
[3] Disponível em: <https://numpy.org/doc/stable/>
[4] Disponível em: <https://pandas.pydata.org/docs/>
[5] Disponível em: <https://seaborn.pydata.org/api.html>

as bibliotecas Keras[6], PmdARIMA[7], Sklearn[8], Statsmodel[9] e XGBoost [10].

3.2 Obtenção dos dados

Os dados relativos aos casos de dengue foram levantados junto à Secretaria Municipal de Saúde de Cascavel - PR e ao setor de Endemias, nos quais atendem aos requisitos de ética quanto à manipulação de dados em arquivos, do parecer 261/2012-CEP, referente ao processo CAAE nº. 10726712.6.000.0107. Além disso, não foram utilizadas informações específicas que possam identificar algum indivíduo conforme a Lei 13.709 de 14 de agosto de 2018. As informações dos pacientes consideradas nesta pesquisa envolvem apenas a data inicial dos primeiros sintomas da doença, referentes ao período disponível dos dados, de 2007 à 2020, sem distinção entre os tipos e sorotipos de dengue.

Além das informações acerca dos casos positivos da doença, fez-se o levantamento dos dados climáticos, do mesmo período e localização. Tais informações envolvem as variáveis: umidade relativa do ar (%) precipitação total (mm) e temperatura média (°C), em granularidade diária. A escolha destas variáveis foi baseada nos estudos de Aburas, Cetiner e Sari (2010), Guo et al. (2017), Mittelmann e Soares (2017a), e Mittelmann e Soares (2017b). Tal conjunto de dados foi obtido junto ao SIMEPAR (Sistema de Tecnologia e Monitoramento Ambiental do Paraná) (SIMEPAR, 2022).

3.2.1 Remoção e tratamento de ruídos

Os ruídos encontrados na base adquirida foram descartados durante a consolidação dos dados. Tais elementos caracterizam-se por datas que não estão contidas no período estudado (janeiro de 2007 a dezembro de 2020), e por classificações finais de dengue que não estão de acordo com a especificação do dicionário de dados.

Além da remoção de amostras datadas fora do período de interesse, fez-se o descarte dos registros de casos suspeitos de dengue não confirmados. A base obtida junto à secretaria de saúde envolvia informações de todos os casos testados no município, uma vez que o protocolo de teste tem notificação compulsória. Visto que o interesse recai apenas sobre os casos positivos, todos os registros em que os resultados dos exames laboratoriais foram negativos, foram removidos.

Na base de dados meterológicos, notou-se que havia dados faltantes referentes a temperatura média dos dias 16 e 17 de janeiro de 2017. Visando preencher tal lacuna na série temporal, foi adotado uma média ponderada dos cinco dias anteriores, para suprir os dados faltantes, com

[6] Disponível em: <https://keras.io/api/>
[7] Disponível em: <https://pypi.org/project/pmdarima/>
[8] Disponível em: <https://scikit-learn.org/0.21/documentation.html>
[9] Disponível em: <https://www.statsmodels.org/stable/user-guide.html>
[10] Disponível em: <https://xgboost.readthedocs.io/en/stable/get_started.html>

base nas Equações 8 e 9 executadas de forma sequencial, de forma que as amostras mais recentes possuam maior influência sobre a data calculada.

$$T_{16/01} = \frac{5 * T_{15/01} + 3 * T_{14/01} + 1.725 * T_{13/01} + 0.25 * T_{12/01} + 0.025 * T_{11/01}}{10} \tag{8}$$

$$T_{17/01} = \frac{5 * T_{16/01} + 3 * T_{15/01} + 1.725 * T_{14/01} + 0.25 * T_{13/01} + 0.025 * T_{12/01}}{10} \tag{9}$$

3.2.2 Tratamento dos dados

O conjunto dos dados climáticos em específico, possui uma amostragem diária. Contudo, para os experimentos realizados foi necessário transformá-los em granularidades semanal e mensal. Para as variáveis referentes à temperatura média e umidade relativa, foi realizada a média aritmética simples dos dados. Em contrapartida, para a transformação da variável referente à precipitação, foi realizado o somatório dos valores para obter a granularidade desejada.

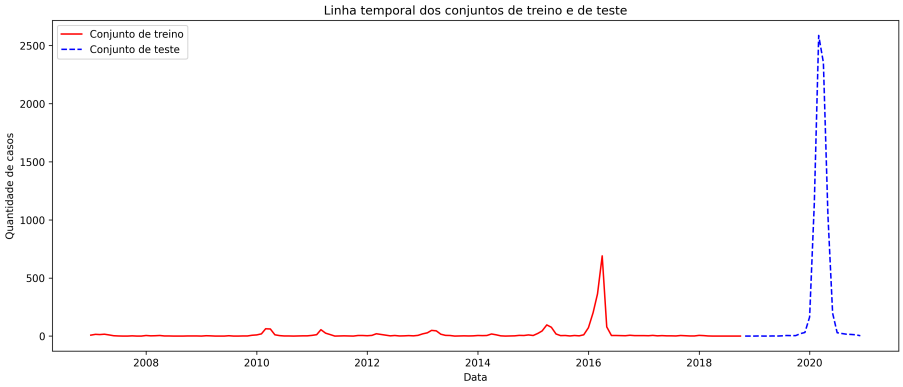
Os dados históricos da dengue foram tratados em dois passos. Primeiramente, agrupou-se o número de casos para cada dia. Se, por ventura, não houvesse casos registrados em uma determinada data, foi atribuído o valor zero. Após este processo, foi realizada a mesma estratégia aplicada à variável climática precipitação: fez-se a contagem de todos os casos ocorridos em períodos semanais e mensais.

A randomização das amostras não foi realizada para manter a série histórica e a sincronia no qual os dados são correlatos, uma vez que considera-se haver dependência temporal entre as amostras.

3.3 Divisão dos dados

Após a consolidação dos dados, foi realizada a divisão da base em dois grupos: conjunto de treino e conjunto de teste. O primeiro possui o objetivo de servir de base para o aprendizado e calibração dos modelos escolhidos, a fim de estimar os casos de dengue. Já o conjunto de teste, tem como objetivo servir de critério para a avaliação e análise dos modelos calibrados. Tais conjuntos correspondem à 85% e 15%, respectivamente, da base de dados consolidada. Tal divisão foi baseada nos trabalhos de Pham et al. (2015) e Azhar, Marina e Anwar (2017). Os conjuntos resultam no período de janeiro de 2007 a outubro de 2018, na composição do grupo de treino, e o período de novembro de 2018 a dezembro de 2020, na composição do grupo de teste, assim como ilustrado na Figura 8, indicados pelas linhas vermelha (contínua) e azul (tracejada), respectivamente. A informação representada corresponde ao número de casos positivos observados ao longo dos quatorze anos.

Figura 8 – Divisão temporal entre os conjuntos de treino e teste



Fonte: o autor

3.4 Treinamento e avaliação dos Modelos

Os modelos matemáticos e estatísticos implementados foram a Média Móvel (MM), a Suavização Exponencial (SE) e o ARIMA. Os modelos de aprendizagem de máquina foram o *Support Vector Machine* (SVM), *Multilayer Perceptron* (MLP), *Random Forest* (RF), *Recurrent Neural Network* (RNN) baseado em estratégias LSTM e GRU, *Bagging* usando os classificadores DT, SVM e MLP, Floresta Aleatória (RF) e *Extreme Boosting* usando as abordagens Linear e Hierárquica. As abordagens citadas foram calibradas de acordo com seus respectivos parâmetros, utilizando *grid search* orientado pela raiz do erro médio quadrático. Por fim, as técnicas foram avaliadas através das métricas do Erro Médio Absoluto (MAE), Erro Médio Quadrático (MSE) e Raíz do Erro Médio Quadrático (RMSE).

3.5 Parâmetros dos modelos preditivos

Nesta seção são apresentadas as configurações testadas para cada um dos modelos implementados na seção 3.4. As configurações que levaram aos melhores desempenhos serão apresentadas no Capítulo 4.

Para o modelo da média móvel, o único parâmetro a ser testado com diferentes valores é o número de amostras n considerado para o cálculo da média, em que $n \in [3, 19] \; \forall \; n \in \mathbb{N}$

A implementação do modelo de suavização exponencial simples, fornecida pela biblioteca *Statsmodel*, permite a especificação do nível de suavização do modelo. Tal parâmetro foi testado com os valores do conjunto $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

O modelo ARIMA conta com três principais parâmetros a serem testados, os quais p, d e q, que correspondem à ordem do modelo auto regressivo, ao grau de diferenciação e à ordem da média móvel, respectivamente. Os valores testados estão representados na Tabela 1.

Tabela 1 – Parâmetros do modelo ARIMA avaliados no *GridSearch*

ARIMA	
Parâmetro	**Valores testados**
p	{0, 1, 2, 3, 4, 5}
d	{0, 1, 2, 3, 4, 5}
q	{3, 4, 5, 6, 7, 8, 9, 10}

Ao utilizar a estrutura do SVM, implementada pela biblioteca Scikit-learn, pode-se fazer uso de vários parâmetros a fim de criar modelos bastante diversos. Neste estudo, foram variados os parâmetros correspondentes ao tipo de *kernel*, núcleo para a elevação do espaço dimensional, e ao parâmetro regularizador C (custo dos erros). Os valores utilizados na busca pelo melhor modelo estão representados na Tabela 2.

Tabela 2 – Parâmetros do modelo SVM avaliados no *GridSearch*

SVM	
Parâmetro	**Valores testados**
C	{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}
kernel	{*linear,rbf, poly*}

Assim como no SVM, existem vários parâmetros que permitem criar modelos de MLPs bastante diversos. Os parâmetros variados foram os correspondentes ao número de neurônios nas camadas escondidas (*hidden_layer_sizes*), à taxa de aprendizagem (*learning_rate*) e ao número máximo de iterações (*max_iter*) para convergência. O conjunto que contém os valores testados em cada parâmetro estão listados na Tabela 3.

Tabela 3 – Parâmetros do modelo MLP avaliados no *GridSearch*

MLP	
Parâmetro	**Valores testados**
hidden_layer_sizes	{100, 200, 300, 400, 500}
learning_rate	{0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01}
max_iter	{400, 600, 800, 1000}

Ainda utilizando a biblioteca Sckit-learn, a implementação de florestas aleatórias conta com diversos parâmetros para criar seus modelos. Aqueles que foram testados correspondem ao critério de escolha entre as árvores (*criterion*), o número de árvores na floresta (*n_estimators*) e ao número mínimo de amostras nas folhas (*min_samples_leaf*). Os valores variados para cada parâmetro do modelo, são apresentados na Tabela 4.

Para a estratégia da RNN, foram criados 8 modelos para serem testados, utilizando as células LSTM e GRU, trabalhando tanto de forma separada quanto em conjunto. Os modelos em questão estão descritos na Tabela 5, doravante referidos através das siglas representada entre parênteses. Todos eles foram treinados por 100 e por 1000 épocas para aumentar o campo de busca pela melhor configuração.

Tabela 4 – Parâmetros do modelo RF avaliados no *GridSearch*

RF	
Parâmetro	**Valores testados**
criterion	{*squared_error, absolute_error*}
n_estimators	{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500}
min_samples_leaf	{2, 3, 4}

Tabela 5 – Modelos de RNR avaliados durante os experimentos

1 camada LSTM (R1L)
2 camadas LSTM (R2L)
3 camadas LSTM (R3L)
1 camada GRU (R1G)
2 camadas GRU (R2G)
3 camadas GRU (R3G)
1 camada LSTM + 1 camada GRU (R1L1G)
1 camada LSTM + 1 camada GRU + 1 camada LSTM + 1 camada GRU (R1L1G1L1G)

A técnica de *Bagging* permite agregar outros modelos de aprendizagem de máquina para fazer as predições de forma mais robusta. Os modelos básicos utilizados foram o MLP, SVR e a Árvore de Decisão, contendo 100 estimadores. É importante ressaltar que os modelos utilizados dentro do *Bagging* não sofreram alterações em seus parâmetros, pois a intenção foi a de manter esses modelos relativamente "fracos" em prol de construir um modelo mais diverso e robusto utilizando a combinação destes.

O único parâmetro variado (*booster*) para o modelo *Extreme Gradient Boosting* envolve a estrutura em que o *booster* é baseado: em árvores e em funções lineares (Tabela 6). Assim como no *Bagging*, foi mantido um número fixo de 100 estimadores para esta técnica.

Tabela 6 – Parâmetros do modelo XGBoost avaliados no *GridSearch*

XGBoost	
Parâmetro	**Valores testados**
booster	{*gblinear*, *gbtree*}

No próximo capítulo serão apresentadas as melhores configurações adotadas em cada modelo bem como o desempenho alcançado por eles. Além disso será apresentada a análise comparativa dos desempenhos e a definição da melhor abordagem para o nosso problema.

4

Resultados Experimentais

Os dados processados, utilizados e analisados para a obtenção dos resultados aqui descritos correspondem ao período de janeiro de 2007 a dezembro de 2020. Tais informações incluem o número de casos confirmados para dengue e as variáveis climáticas elencadas anteriormente.

Neste capítulo são descritos os resultados da análise da série temporal estudada, avaliando a existência de sazonalidade nos dados levantados. Além disso, é apresentada a análise da correlação entre as variáveis de suporte e a variável alvo do estudo e, por fim, as predições realizadas pelos nove modelos escolhidos. Todos os resultados estão descritos textualmente e graficamente nas próximas seções.

4.1 Distribuição do número de casos de dengue ao longo do período estudado

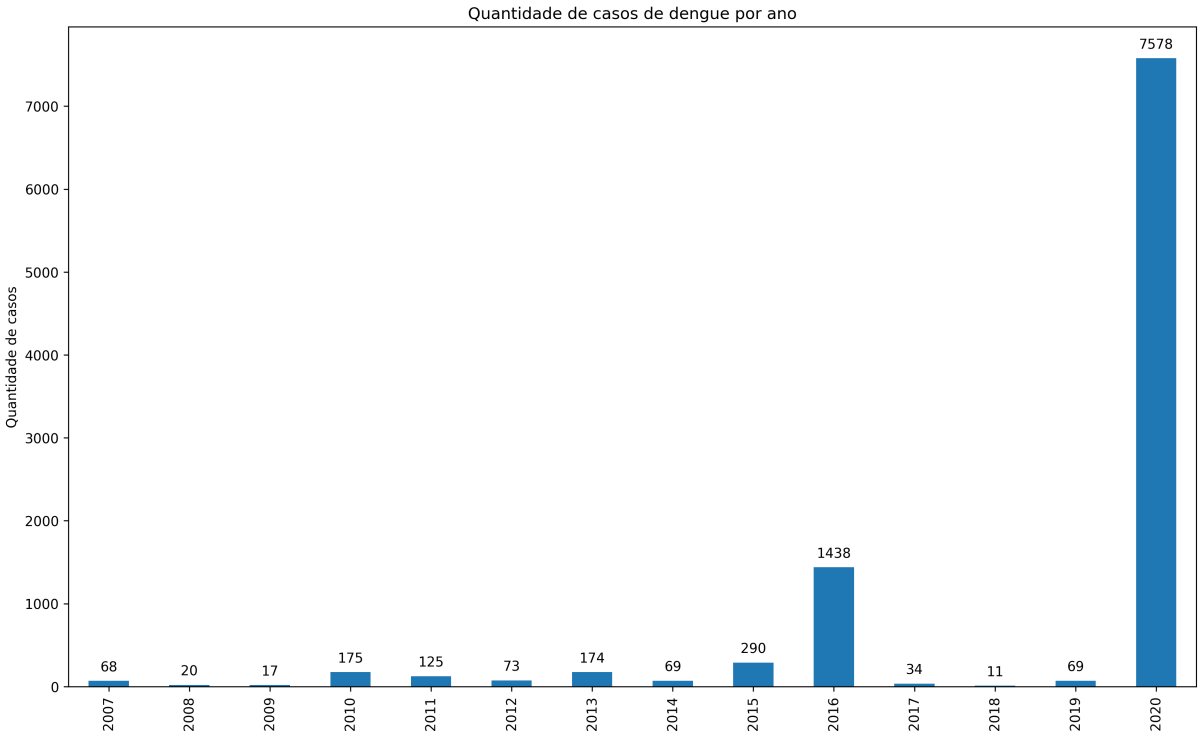
A exploração dos dados disponíveis sobre o fenômeno em questão possibilita a tomada de melhores decisões para o projeto. Dessa forma, foi realizada uma análise exploratória das informações levantadas visando conhecer o contexto da aplicação da pesquisa de forma mais aprofundada.

A filtragem do número de casos de dengue não levou em conta a diferenciação entre sorotipos (DENV-1, DENV-2, DENV-3, DENV-4 e DENV-5), nem considerou casos que seriam estrangeiros, ou seja, casos de pessoas que não residam em Cascavel mas foram registrados na cidade são parte desse conjunto. Essa distinção não foi realizada para manter o termo de compromisso e ética ao obter esses dados.

A Figura 9 exibe a quantidade de casos em cada ano do período estudado. A média anual é de quase 724 confirmados, apresentando um desvio padrão de 2006, o que indica uma variação considerável entre as amostras. Observando-se a curva presente na figura pode-se notar que

em 2016 e 2020 houve surtos de casos de dengue, ultrapassando 1400 ocorrências em 2016 e chegando a mais de 7500 casos em 2020. Outro fato interessante é o controle que houve sobre a doença no período entre os dois surtos, onde soma-se apenas 114 casos positivos entre 2017 e 2019. Na Tabela 7 é possível perceber que 2016 e 2020 têm comportamento bastante distinto do número de casos obsevados no restante do período.

Figura 9 – Quantidade de casos de dengue por ano



Fonte: o autor

Nos gráficos ilustrados nas Figuras 35 à 48 (localizadas no Apêndice A) são apresentados o número de casos positivos diários ao longo dos anos de 2007 até 2020.

Buscando entender os surtos ocorridos nos anos referidos, a Tabela 8 apresenta as médias diárias das informações climáticas e dos casos positivos de dengue ao longo do período estudado. Os valores referentes aos anos de 2016 e 2020 estão destacados na tabela, contudo não observou-se nenhum fenômeno climático que insinuasse um aumento tão acentuado na quantidade de casos registradas. O único fato constatado foi que em 2020 houve a menor incidência de chuva em todo o período analisado, o que acarretou na menor umidade relativa do período.

Ao analisar o número de casos de dengue distribuídos ao longo dos meses em uma representação estatística circular, pode-se confirmar essa intensidade maior em alguns meses. Esta informação é representada na Figura 10. Na distribuição circular dos casos, cada mês corresponde a uma faixa de 30°. Assim, pode-se considerar que cada mês do ano corresponde a uma direção específica. Os maiores índices de casos positivos foram constatados em fevereiro

Tabela 7 – Número de casos positivos de dengue por ano

Ano	**Nº de casos confirmados**
2007	67
2008	20
2009	17
2010	175
2011	125
2012	73
2013	174
2014	64
2015	290
2016	1438
2017	34
2018	11
2019	69
2020	7578

Tabela 8 – Médias diárias das informações climáticas e casos positivos de dengue para o período de 2007 a 2020

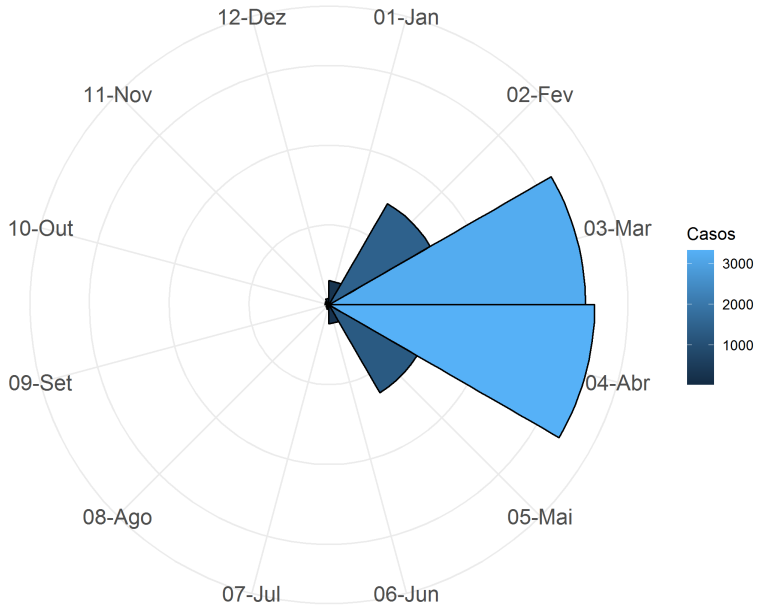
Ano	Precipitacão (mm)	Temp. Média (°C)	Umid. Relativa (%)	Número de Casos
2007	4,0690	20,9215	78,7967	0,1863
2008	4,1287	20,2811	79,8778	0,0546
2009	5,2595	20,5424	76,4320	0,0466
2010	4,7510	20,1119	73,5971	0,4795
2011	4,6115	20,1275	74,6581	0,3425
2012	5,2350	20,8057	73,1776	0,1995
2013	5,8768	19,8503	76,2472	0,4807
2014	6,3852	20,7009	77,0122	0,1890
2016	5,9825	19,9609	77,2205	3,9290
2017	8,1337	20,2713	73,0464	0,0932
2018	5,5129	19,6375	73,8880	0,0301
2019	3,9573	20,4514	71,7904	0,1890
2020	3,1399	20,3273	67,3603	20,7049

(1463), março (3211), abril (3326) e maio (1270). Por outro lado, nos meses de julho (58), agosto (42), setembro (35), outubro (50), novembro (56) e dezembro (83), a frequência de ocorrências de dengue foi bem menor.

Esse fenômeno está relacionado diretamente às características do clima subtropical da cidade de Cascavel, no qual costuma-se observar uma incidência maior de chuvas no verão e no outono em comparação ao inverno. Por consequência, a probabilidade de acúmulo de água parada para criadouros do mosquito da dengue aumenta, além das temperaturas mais altas que também contribuem para a proliferação do vetor.

Na Figura 10 é possível perceber uma frequência maior de ocorrências de dengue no

Figura 10 – Representação circular da distribuição mensal dos casos de dengue



Fonte: o autor

primeiro semestre do ano (principalmente no período de fevereiro a maio). Para entender a proporção do número de casos positivos é apresentado na Figura 11, um gráfico de pizza com o número de ocorrências separados por mês. Na ilustração cada faixa corresponde a um mês específico. Nas legendas, são apresentados o mês de referência, o número de casos ocorridos naquele período e o percentual perante o total de casos ao longo dos doze meses, respectivamente.

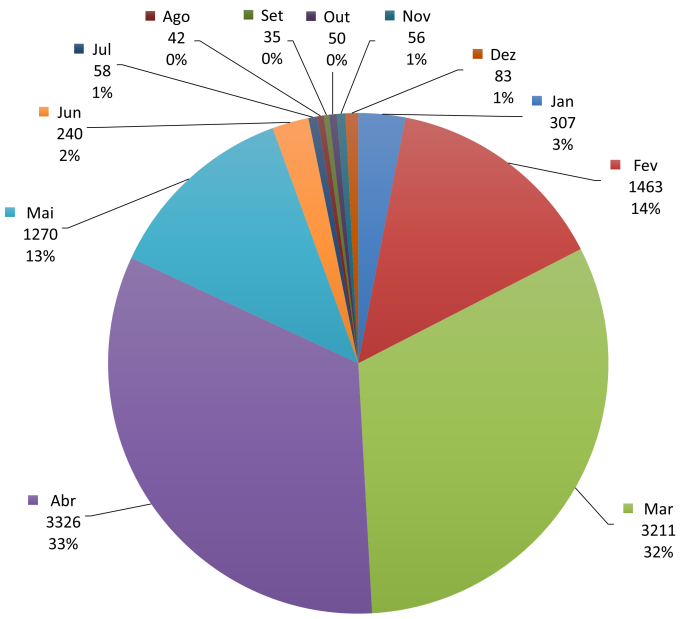
De acordo com a estatística circular da Figura 10, os meses de fevereiro, março, abril e maio compreendem uma faixa extensa do gráfico da Figura 11. O total de ocorrências observadas nesses quatro meses corresponde a 92% do total levantado, indicando uma forte tendência no comportamento da doença no município de Cascavel.

Visando avaliar se há um comportamento homogêneo na distribuição dos casos da doença ao longo do ano foi realizado o teste de periodicidade de Rayleigh (BRAZIER, 1994). Uma probabilidade menor que o nível de significância escolhido indica que os dados não têm uma distribuição uniforme (rejeita-se a hipótese nula que indica homogeneidade na distribuição dos dados). Neste caso, haverá evidências de que a distribuição é mais concentrada em determinadas direções.

O valor obtido pelo teste, $P - value = 0$, indica que a distribuição dos casos de dengue não segue um comportamento regular, considerando uma significância de 5%. Tal fato implica na existência de certa sazonalidade nos casos de dengue.

Visando avaliar se há uma concentração maior em alguma época específica do ano, foi

Figura 11 – Proporção do número de casos de dengue para cada mês



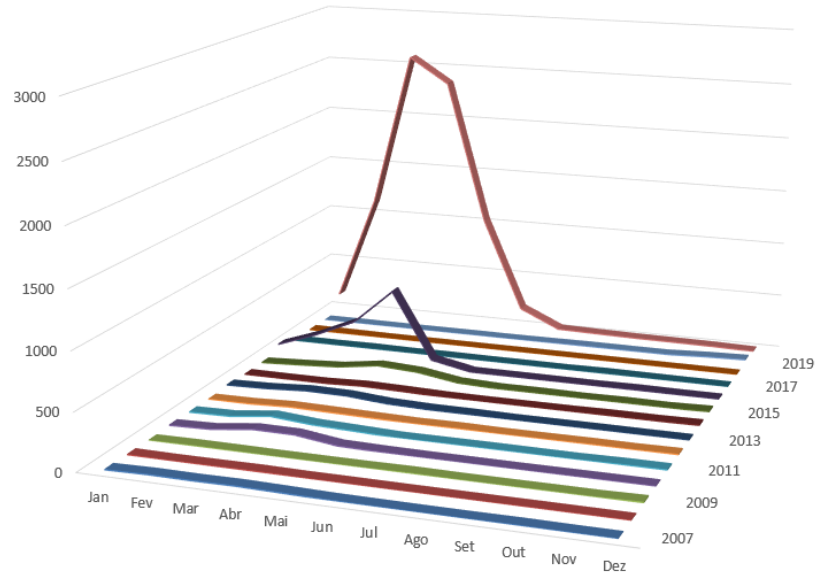
Fonte: o autor

calculado o vetor circular médio P, em que $P \in [0, 1]$, cuja intensidade é usada para determinar o quão forte é a média da distribuição circular. Um valor igual a 1 indica que todas as amostras seguem a mesma direção implicando que todos os casos teriam ocorrido no mesmo mês, enquanto um valor próximo de zero indica que não há uma direção de maior dominância.

Analisando a intensidade do vetor médio obtido, $P = 0,8089$, percebe-se que há uma direção específica que envolve grande parte da distribuição dos casos de dengue. Tal fato pode ser constatado na Figura 10 onde observa-se um pico que se inicia em janeiro e se conclui em junho.

Considerando a existência de sazonalidade nos dados (tendência na direção leste conforme a Figura 10), foi realizada uma análise na distribuição dos casos positivos ao longo dos 14 anos para saber se essa tendência é recorrente ou se foi causada pela concentração dos casos observados nos anos de 2016 e 2020 que, dada sua magnitude, poderiam criar tendências. A Figura 12 apresenta o comportamento das distribuições para todos os anos (representados no eixo Z) durante cada um de seus meses (eixo X). Já no eixo vertical temos a quantidade de casos observados para cada coordenada [ano,mês]. Uma representação alternativa é apresentada na Figura 13. Na ilustração são detalhados os valores de casos confirmados em todos os meses observados entre os anos de 2007 e 2020. É possível notar que os maiores valores (nas cores mais claras) estão compreendidos entre os meses de fevereiro e maio ao longo dos anos, confirmando a sazonalidade observada anteriormente.

Figura 12 – Números de casos positivos de dengue distribuídos ao longo dos meses do período de 2007 a 2020



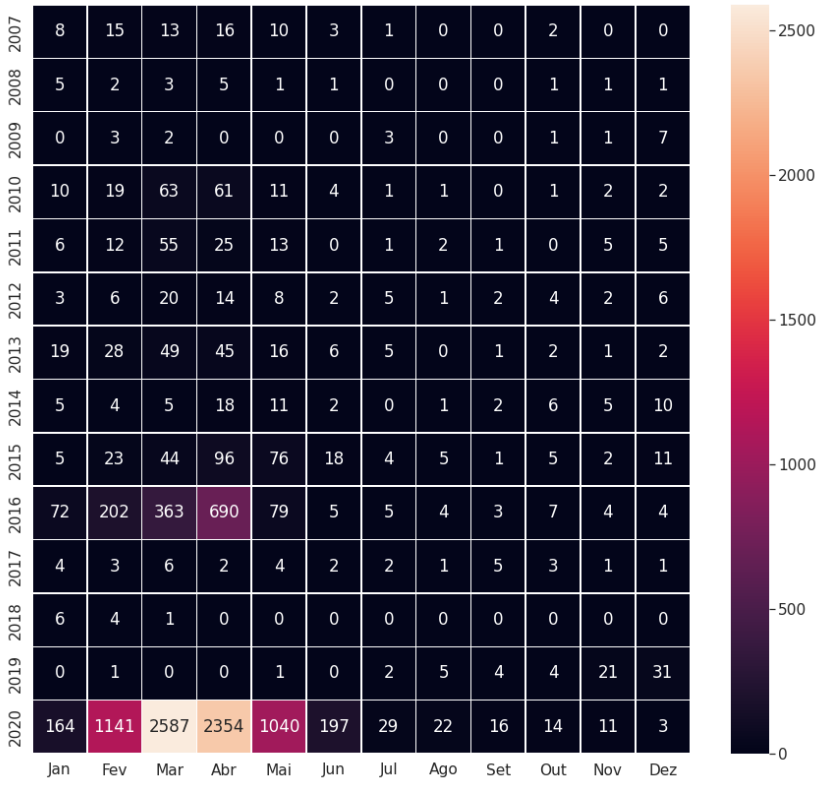
Fonte: o autor

4.2 Estudo de correlação entre as variáveis climáticas e casos de dengue

Buscou-se analisar a relação entre as variáveis climáticas (umidade relativa do ar, precipitação e temperatura média) perante a quantidade de casos positivos ao longo de todo o período, com amostragem diária, primeiramente. Para tal fim, utilizou-se o coeficiente de correlação de Pearson, que está representado na forma de matriz de correlação na Figura 14. Entretanto, as correlações entre as variáveis climáticas e a quantidade de casos (última linha e última coluna) não mostrou uma intensidade significativa, a ponto de não superar um escalar de 0,01.

Levando em consideração o tempo de vida do mosquito, em torno de 20 a 30 dias para tornar-se adulto, infectar-se com o vírus e assim ser capaz de transmitir a doença, decidiu-se por deslocar as variáveis climáticas em retrocesso, isso é, acoplar a quantidade de casos de dengue com os dados climáticos de 20 a 30 dias atrás. O objetivo é avaliar que, se os fatores climáticos influenciam no ciclo de vida do mosquito, com a incidência de chuvas e temperaturas altas, há uma maior eclosão dos ovos e dispara-se o início do ciclo. Dessa forma, optou-se por analisar essa relação considerando todo o possível ciclo de vida do *Aedes*, haja visto que não há uma concordância quanto à extensão do prazo entre a eclosão do ovo e o momento em que o mosquito se torna apto ao contágio. Os resultados desses experimentos estão ilustrados na forma de matriz de correlação nas Figuras 49 à 58, localizadas no Apêndice B.

Figura 13 – Mapa de calor indicando a distribuição dos casos confirmados durante todos os 168 meses estudados



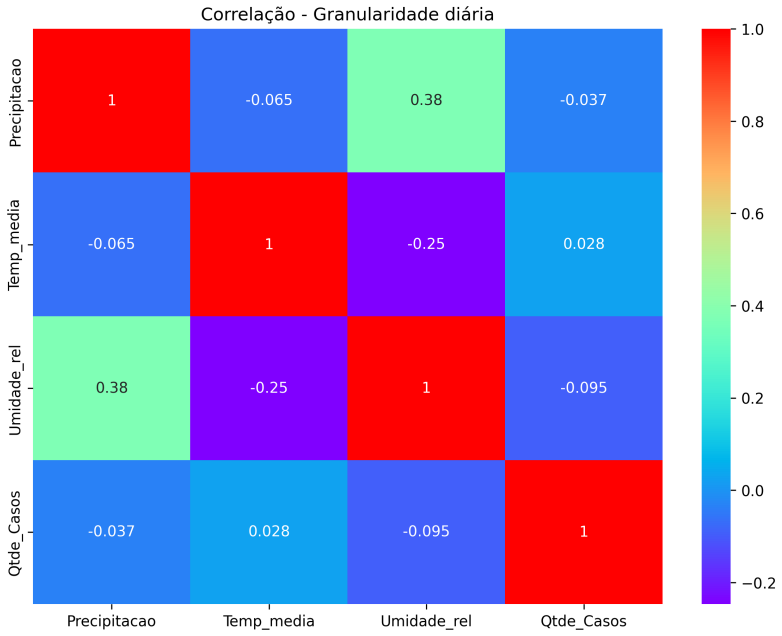
Fonte: o autor

Ao observar os valores dos coeficientes apresentados, obteve-se uma certa melhora na correlação entre as variáveis estudadas. Os valores observados, no entanto, mostraram-se bastante reduzidos. Uma forma de identificar uma relação mais representativa foi agrupar os dados climáticos e dos casos positivos em amostragens mensais. Os níveis da correlação resultante dessa análise são exibidos na Figura 15. Percebe-se que os valores realmente mostraram um aumento na correlação entre as variáveis de estudo.

Na Figura 15 entretanto, não foi considerado o deslocamento temporal de trinta dias necessários para que o vetor esteja apto à transmissão. Então, foi realizado essa manipulação das variáveis climáticas, retrocedendo o período de um mês, de forma que, se em um determinado período houve altas temperaturas e bastante chuva, a tendência é que seu impacto seja percebido no mês seguinte, quando os mosquitos tornam-se capazes da transmissão da dengue. Os valores obtidos por esta correlação são apresentados na Figura 16.

Em comparação às correlações anteriores, os valores agrupados mensalmente com deslocamento de um mês frente às variáveis climáticas, mostraram coeficientes de correlação mais interessantes, obtendo os maiores valores. Isso comprova que os fatores climáticos ocorridos em um mês terão algum impacto (ainda que fraco) no mês seguinte, visto a duração de todo o

Figura 14 – Matriz de correlação com granularidade diária (sem deslocamento)



Fonte: o autor

ciclo do vetor.

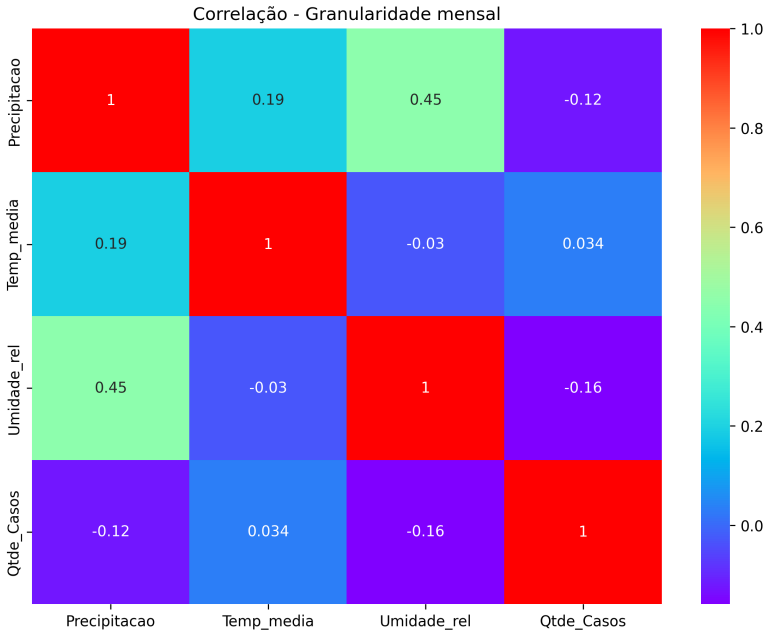
Os experimentos apresentados anteriormente foram fundamentais pois, sabendo a forma mais eficiente de trabalhar os dados, foi possível construir um modelo de predição mais eficaz e com resultados mais precisos. Desse modo, definiu-se por utilizar a granularidade mensal nos dados nas etapas seguintes.

4.3 Configuração resultante de cada modelo

Com a granularidade mensal definida e com o aspecto do retrocesso das variáveis climáticas em relação ao número de casos, realizou-se a modelagem das nove técnicas propostas para a predição de casos de dengue.

Após a execução do *grid search* para cada modelo com seus parâmetros específicos, descritos no Capítulo 3 na seção 3.5, obteve-se a configuração mais adequada dos modelos cujos parâmetros são detalhados na Tabela 9. Tais valores são aqueles que levaram o modelo a obter a menor raiz do erro quadrático médio. Essas configurações foram as adotadas na avaliação dos modelos sobre o conjunto de teste, cujos resultados são apresentados na seção a seguir.

Figura 15 – Matriz de confusão com dados agrupados mensalmente



Fonte: o autor

4.4 Resultados dos modelos preditivos

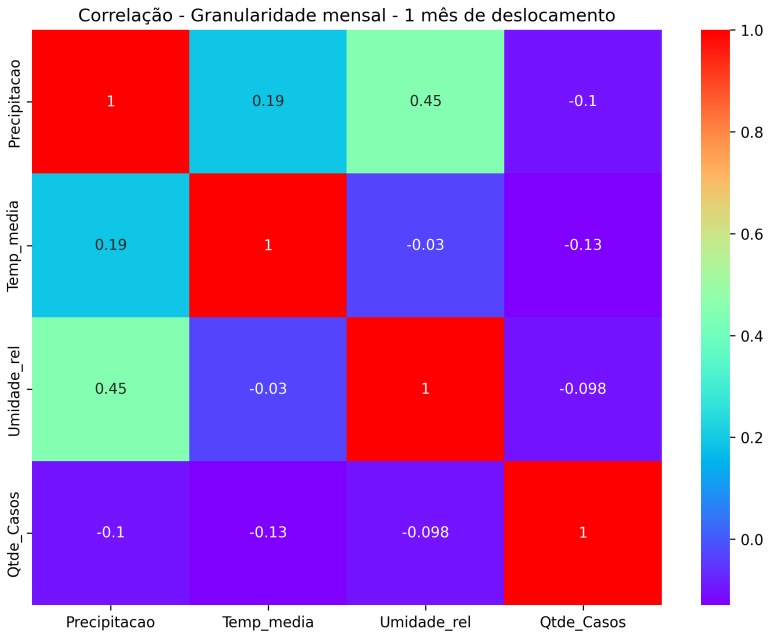
Após a devida calibração dos modelos foi realizada a predição do número de casos de dengue no período de novembro de 2018 a dezembro de 2020 (correspondente ao conjunto de teste) e a posterior comparação com os dados reais observados para o período. Para avaliar os resultados obtidos foram empregados o erro médio absoluto (MAE), o erro médio quadrático (MSE) e a raíz do erro médio quadrático (RMSE).

Ao testar as predições realizadas pelos modelos estatísticos, a média móvel obteve a melhor performance dentre as três técnicas, apresentando um erro médio absoluto de 192,962, erro médio quadrático de 166.552,321 e raiz do erro médio quadrático de 408,108. Os resultados obtidos dos modelos estatísticos são apresentados na Tabela 10. O melhor resultado para as abordagens está destacado na tabela.

A análise visual da predição realizada pelos modelos pode contribuir na percepção de seus desempenhos. O comportamento da Média Móvel é apresentado na Figura 17, enquanto as Figuras 18 e 19 ilustram os comportamentos apresentados pelos métodos de Suavização Exponencial e ARIMA, respectivamente.

Nas figuras, adotou-se o padrão para a comparação de modelos sozinhos, no qual a curva na cor vermelha corresponde aos dados reais observados durante o ano e a curva pontilhada com marcadores na cor azul refere-se ao número de casos de dengue estimados pelo modelo em questão. Figuras que não comparam modelos sozinhos serão detalhadas assim que forem citadas.

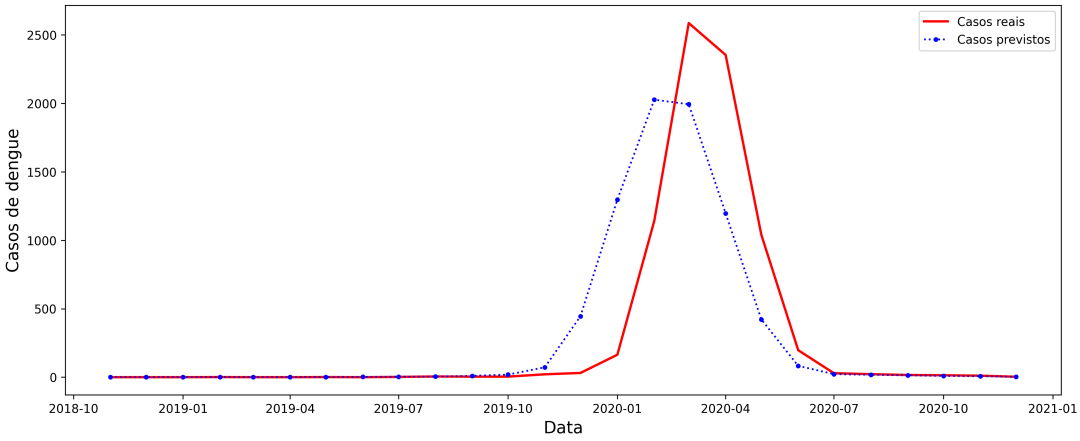
Figura 16 – Matriz de confusão com dados agrupados mensalmente e quantidade de casos deslocada em 1 mês



Fonte: o autor

Observando as Figuras 17 a 19 nota-se que a média móvel manteve sua curva de predição próxima à curva dos dados reais, acompanhando sua maior ascensão e declínio, porém com um pouco de antecedência e não tão intensa quanto o ocorrido realmente.

Figura 17 – Previsão de casos de dengue do modelo média móvel (MM)



Fonte: o autor

A suavização exponencial, todavia, considerou que todos os valores preditos seriam iguais, criando assim uma reta de tendência dos casos de dengue ao longo dos vinte e seis meses de teste. O número de casos estimados pela suavização exponencial para os meses de teste foi

Tabela 9 – Parâmetros selecionados pelo *Grid search*

Método	**Parâmetros**	**Melhor valor**
Média Móvel	n	3
Suavização Exponencial	α	0.1
ARIMA	p	0
	d	1
	q	1
SVM	*kernel*	*linear*
	C	0.1
MLP	*hidden_layer_size*	300
	learning_rate	0.005
	max_iter	800
RNN	Épocas	1000
	Configuração	1 camada LSTM e 1 camada GRU
RF	*criterion*	*squared_error*
	n_estimators	10
	min_samples_leaf	2
Bagging	Modelo	MLP
XGBoost	*Booster*	*Tree*

Tabela 10 – Métricas de precisão alcançadas pelos modelos estatísticos

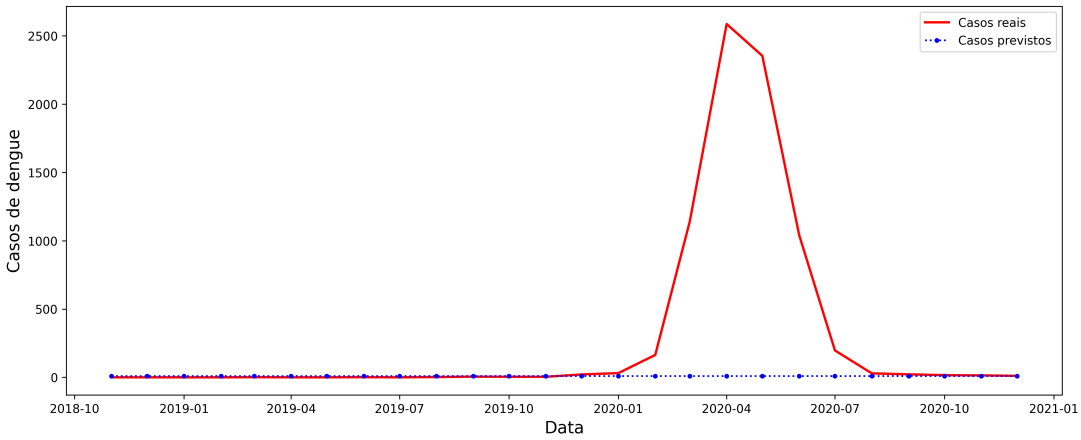
	MM	**SE**	**ARIMA**
MAE	192,962	292,577	299,835
MSE	166.552,321	560.124,900	508.009,355
RMSE	408,108	748,415	712,748

inferior a sete (6,618746), o que contribuiu para os valores tão altos para as métricas de erro da suavização exponencial, uma vez que o número de casos chegou a 2587 em fevereiro de 2020.

O modelo ARIMA por sua vez, previu uma alta no número de casos maior do que realmente aconteceu no começo de 2019 e menor do que ocorrido em 2020. Como analisado na seção 4.1, o modelo ARIMA previu a alta do número de casos no período correto do ano, de janeiro à maio, contudo, houve grande disparidade entre os anos que compõem o período de teste de predição, onde em um ano houve pouquíssimos casos e no seguinte, uma epidemia.

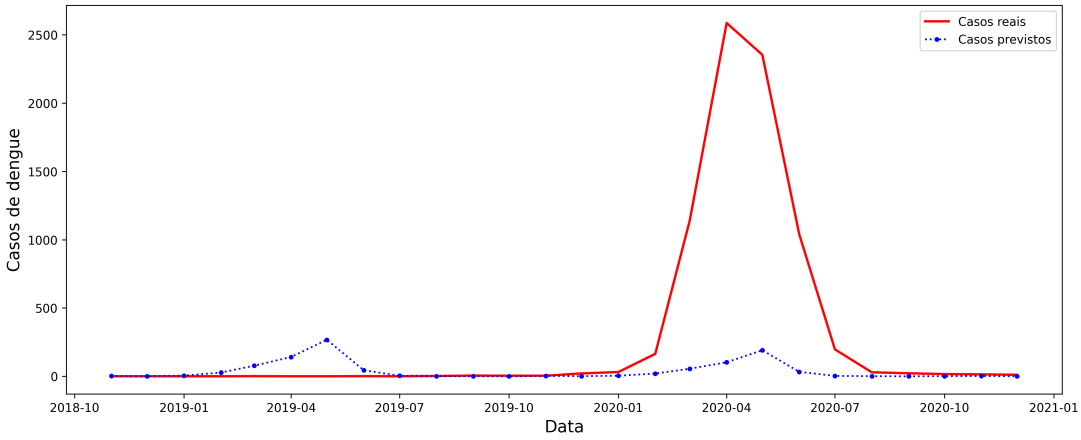
A diferença no comportamento das técnicas estatísticas pode ser percebida quando ilustradas todas as suas curvas de estimação em paralelo à curva real de casos positivos, conforme apresentado na Figura 20, onde a linha azul representa a curva referente à média móvel, a linha verde é referente à suavização exponencial e a linha amarela refere-se ao modelo ARIMA. O desempenho dos métodos em termos numéricos apresentados na Tabela 10 são perceptíveis ao analisarmos a representação visual do comportamento de cada modelo. No gráfico fica evidente que a média móvel se adaptou melhor aos dados, apresentando desempenho mais preciso. A suavização exponencial e o modelo ARIMA alcançaram precisão muito similares, uma a outra, entretanto, o valor do RMSE para o ARIMA lhe dá uma pequena vantagem frente à suavização

Figura 18 – Previsão de casos de dengue do modelo suavização exponencial (SE)



Fonte: o autor

Figura 19 – Previsão de casos de dengue do modelo ARIMA



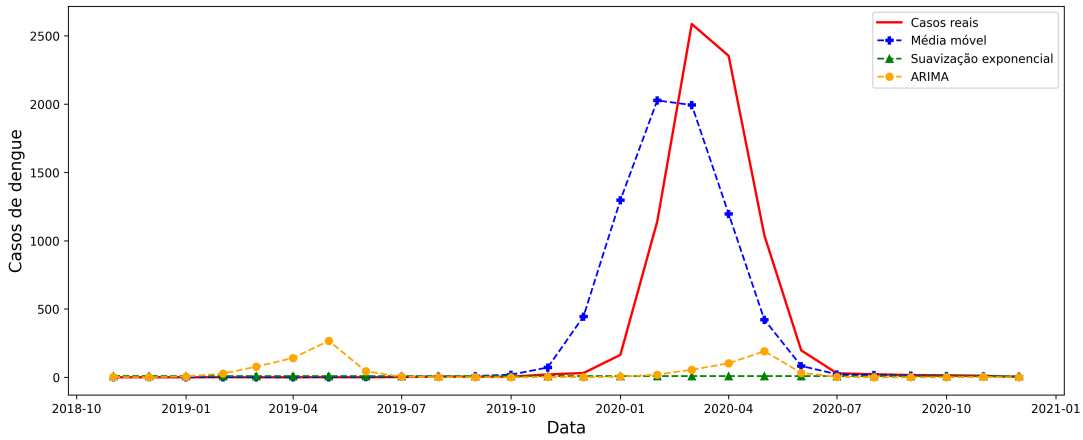
Fonte: o autor

exponencial.

A análise das técnicas que realizam a estimação do número de casos de dengue a partir de conceitos de Aprendizagem de Máquina será dividida em dois grupos. No primeiro serão analisados os regressores que se baseiam em um modelo monolítico (um único regressor) e o segundo grupo compreende as técnicas que se utilizam de conjuntos de modelos para tentar alcançar taxas de precisão melhores.

Dentre os oito modelos de RNN criados, a configuração com uma camada LSTM e uma camada GRU (R1L1G) obteve os melhores valores. Porém, analisando as dez abordagens monolíticas testadas, o *Support Vector Machine* alcançou as melhores métricas. O modelo apresentou um MAE de 0,091 enquanto seu MSE foi de 0,009 com um RMSE de 0,093. Os

Figura 20 – Previsão de casos de dengue dos modelos estatísticos



Fonte: o autor

resultados obtidos pelos modelos monolíticos de aprendizagem de máquina são apresentados na Tabela 11. Em destaque o modelo de melhor desempenho.

Tabela 11 – Métricas apresentadas pelos modelos monolíticos de aprendizagem de máquina

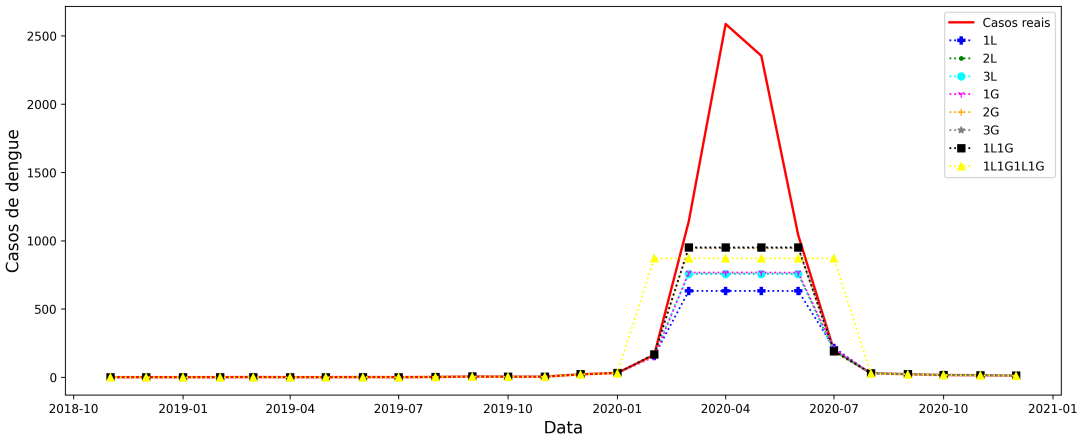
Método	**MAE**	**MSE**	**RMSE**
MLP	68,985	6.569,070	81,050
SVM	0,091	0,009	0,094
R1L	176,755	273.878,453	523,334
R2L	158,331	235.777,279	485,569
R3L	205,209	271.356,907	520,919
R1G	158,280	234.519,266	484,272
R2G	130,108	181.630,465	426,181
R3G	129,595	182.082,114	426,711
R1L1G	128,828	180.330,756	424,654
R1L1G1L1G	128,764	181.093,848	425,551

A razão da escolha do R1L1G como melhor modelo de rede neural recorrente (RNN) pode ser entendida ao se analisar a Figura 21. Na ilustração são representados os comportamentos das oito abordagens implementadas. É possível notar que os modelos geraram predições muito parecidas entre si. Contudo, dentre eles, o modelo que apresentou maior similaridade à curva real foi o R1L1G, composto por uma camada LSTM e uma camada GRU.

O desempenho dos métodos SVM, MLP e do melhor modelo RNN são apresentados nas Figuras 22, 23 e 24, respectivamente. Assim como nas figuras anteriores que comparam modelos sozinhos, a curva na cor vermelha corresponde aos dados reais observados durante o ano. Já a curva pontilhada na cor azul representa os valores estimados pelo modelo de regressão.

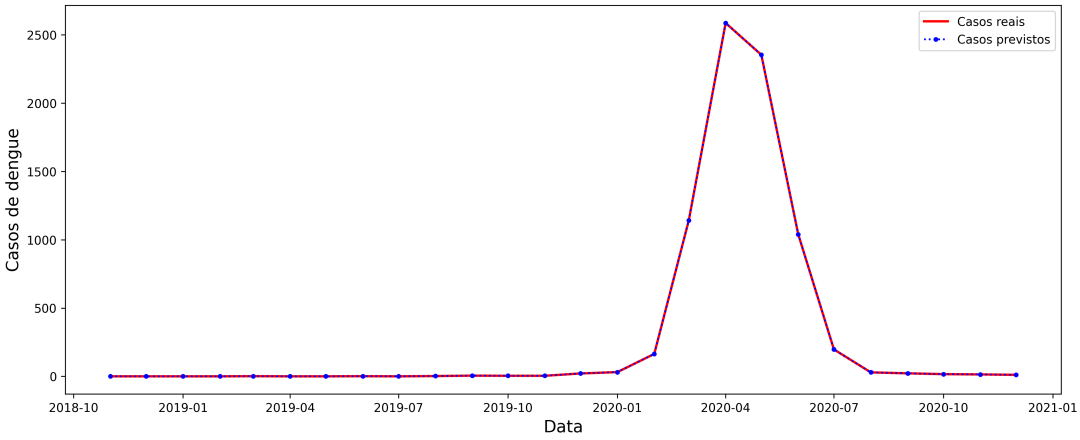
Comparando as curvas de predição construídas para cada modelo nota-se que a curva do SVM (Figura 22) está praticamente sobreposta à curva real, escondendo-a. Isso mostra que

Figura 21 – Previsão de casos de dengue dos modelos RNN



Fonte: o autor

Figura 22 – Previsão de casos de dengue do modelo SVM

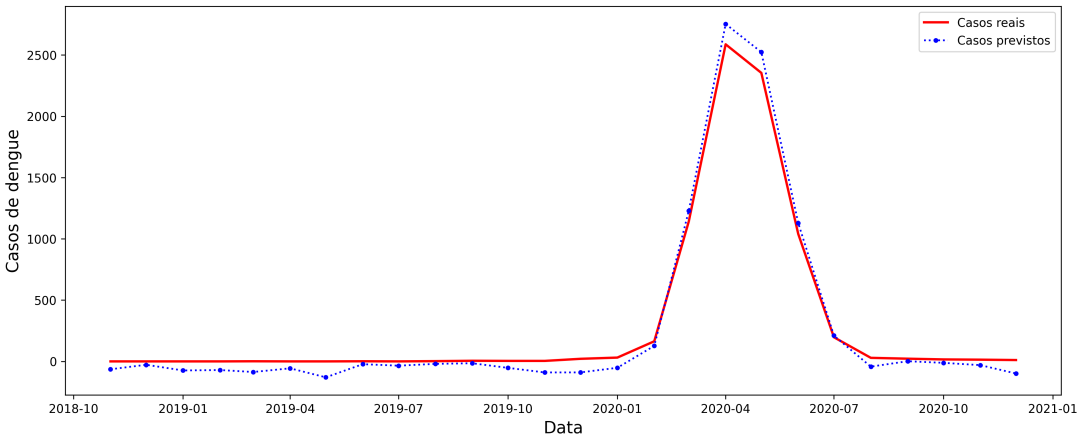


Fonte: o autor

o desempenho deste modelo para a predição foi extremamente preciso. É possível perceber que o MLP (Figura 23), apesar de não ter tido os melhores resultados, apresentou desempenho bastante similar ao SVM, quando analisado suas curvas de predição. O *multilayer perceptron* foi o único regressor que extrapolou o número de casos observados em 2020. Por fim, a rede neural recorrente não obteve desempenho tão bom quanto os demais. Contudo, seu gráfico (Figura 24) mostra que foi coerente com a série real, conseguindo prever o aumento e o declínio no número de casos no período correto, não estimando, no entanto, o surto no ano de 2020 com a intensidade real observada.

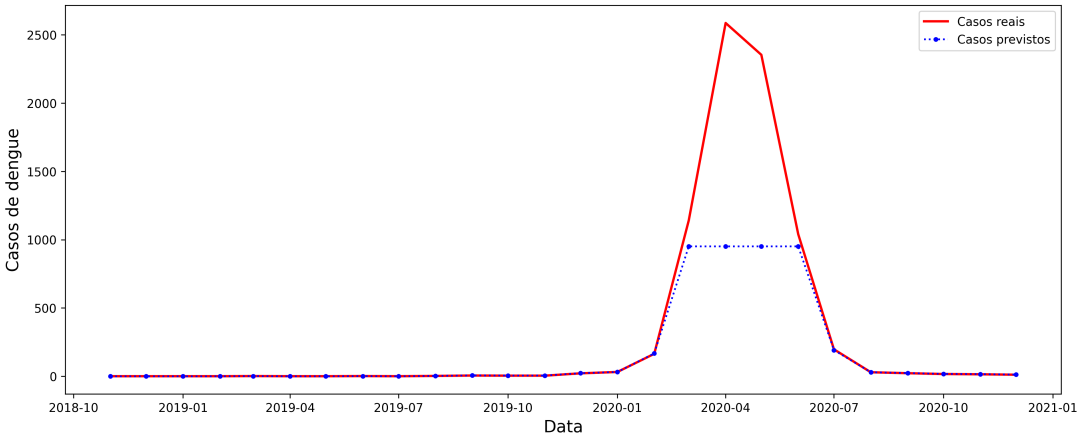
A Figura 25 compara simultaneamente os três modelos monolíticos de aprendizagem de máquina, onde é possível notar o desempenho praticamente sobreposto entre o SVM (destacado pela linha esverdeada) e a curva real. Nota-se também a pequena distinção desta perante o

Figura 23 – Previsão de casos de dengue do modelo MLP



Fonte: o autor

Figura 24 – Previsão de casos de dengue do modelo RNN



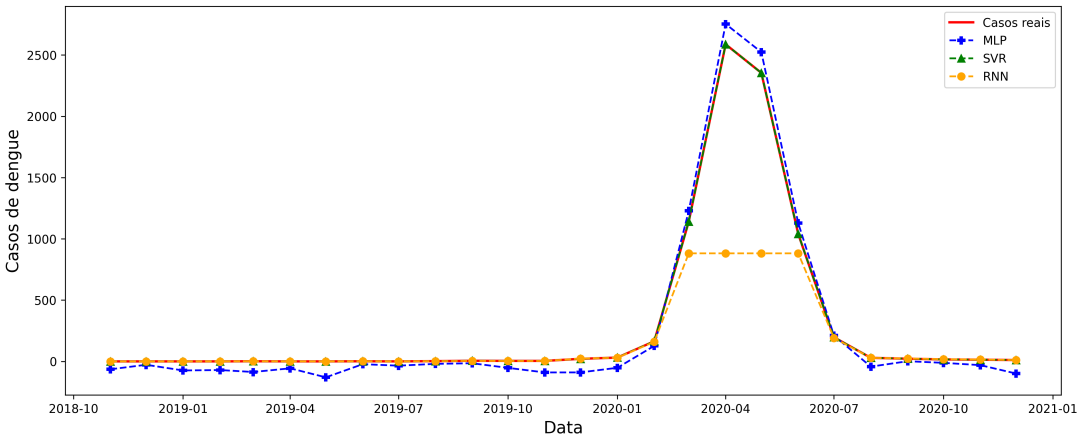
Fonte: o autor

comportamento do MLP (curva azul) e uma discrepância maior em relação ao RNN (destacado em amarelo).

Visando construir modelos de estimação mais robustos e que pudessem contribuir sobre os resultados preditos, foram adotados três estratégias que envolvem uma abordagem em forma de *pool* de modelos regressores, ou seja, utiliza várias instâncias de um modelo para realizar a predição. Tais abordagens são a Floresta Aleatória (RF), o *Bagging* e o *Extreme Gradient Boosting* (XGBoost).

Dentre os modelos *Bagging* utilizados, o regressor que obteve as melhores métricas foi o MLP, e dentre os modelos XGBoost, o melhor *booster* foi o baseado em árvores. Tais comparações são ilustradas nas Figuras 26 e 27, respectivamente. Na primeira figura é possível

Figura 25 – Previsão de casos de dengue dos modelos monolíticos de aprendizagem de máquina

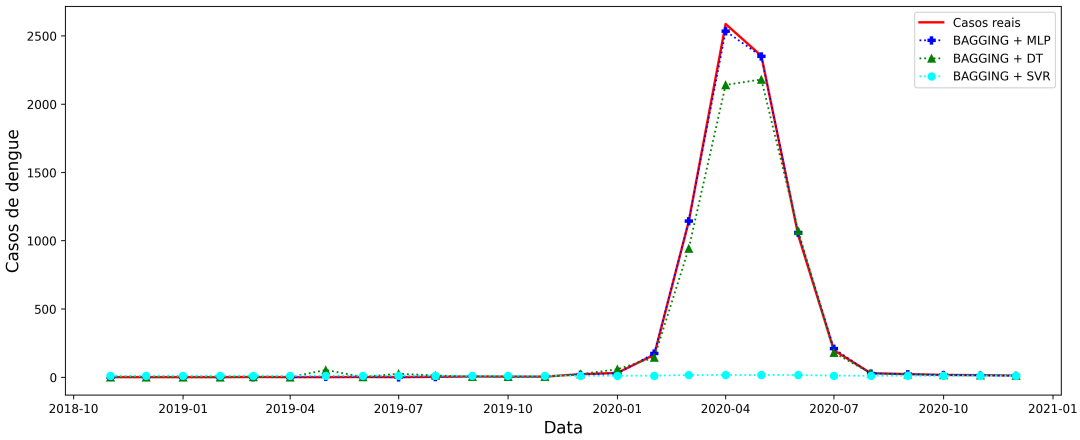


Fonte: o autor

notar que a curva em azul escuro (+) é aquela que apresenta o melhor desempenho dentre as estratégias do *Bagging*. O desempenho da abordagem baseada em árvores de decisão não se mostrou tão interessante, como visto na curva em verde (△). Já a estratégia que combina o Bagging com o SVM apresentou desempenho muito inferior (curva em azul claro sinalizada por ○).

O desempenho dos métodos baseados em XGBoost (Figura 27) mostraram comportamento bastante parecido com os valores reais, tanto para a estratégia linear (curva em azul sinalizada por +) quanto para a adoção de árvores (curva em amarelo sinalizada por △). A melhor alternativa foi definida pelo menor valor de RMSE.

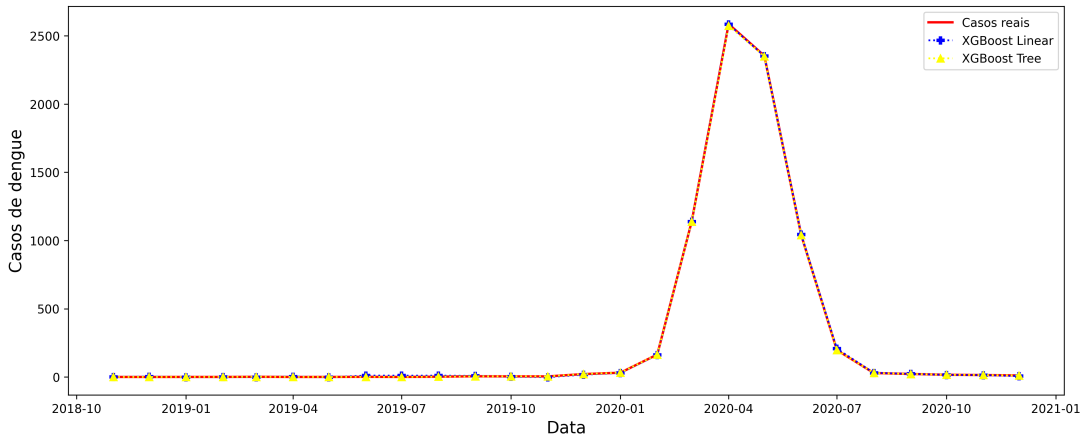
Figura 26 – Previsão de casos de dengue dos modelos *Bagging*



Fonte: o autor

A avaliação das abordagens de regressores múltiplos indicou que o XGBoost apresentou

Figura 27 – Previsão de casos de dengue dos modelos XGBoost



Fonte: o autor

as melhores métricas: MAE igual a 0,798, MSE de 6,740 e RMSE equivalente a 2,596. Na Tabela 12 são apresentadas as métricas obtidas pelas estratégias utilizadas, com destaque para o regressor de melhor desempenho.

Tabela 12 – Métricas apresentadas pelos modelos de aprendizagem de máquina - técnicas de agrupamento

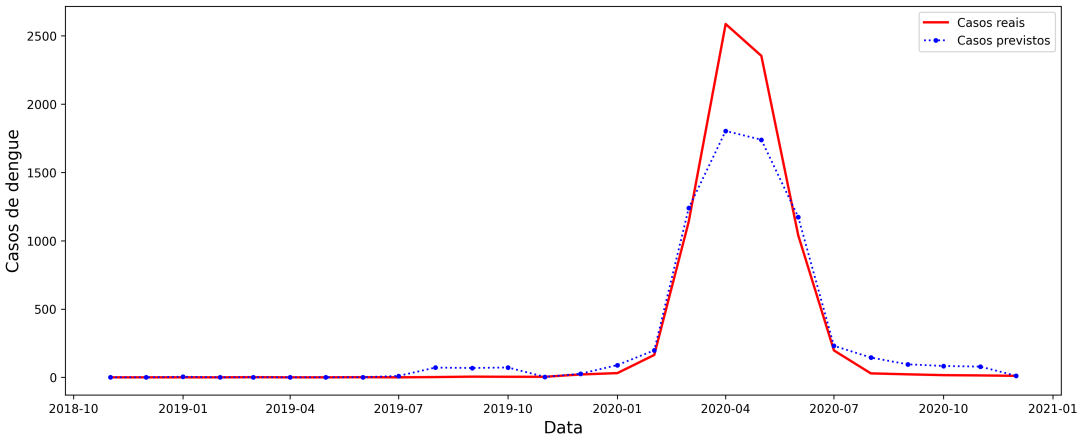
Método	**MAE**	**MSE**	**RMSE**
RF	88,475	40.955,356	202,374
Bagging - MLP	4,745	129,628	11,385
Bagging - SVM	291,748	556.011,662	745,662
Bagging - DT	38,828	10.561,315	102,768
XGBoost - Linear	3,485	19,881	4,459
XGBoost - Trees	0,798	6,740	2,596

Nas Figuras 28, 29 e 30 são ilustrados os comportamentos das predições das abordagens de Floresta Aleatória, *Bagging* (melhor configuração) e XGBoost (melhor configuração), respectivamente.

Ao comparar as três abordagens que utilizam da técnica de *pool* de modelos regressores, ilustrada na Figura 31, nota-se que as curvas dos últimos dois modelos são bastante próximas, distinguindo entre si apenas no mês de março e abril de 2020. A curva da floresta aleatória conseguiu acompanhar a ascensão de casos do ano de 2020, porém não tão fiel quanto a realidade. Tal fato fez com que suas métricas de erro fossem mais elevadas.

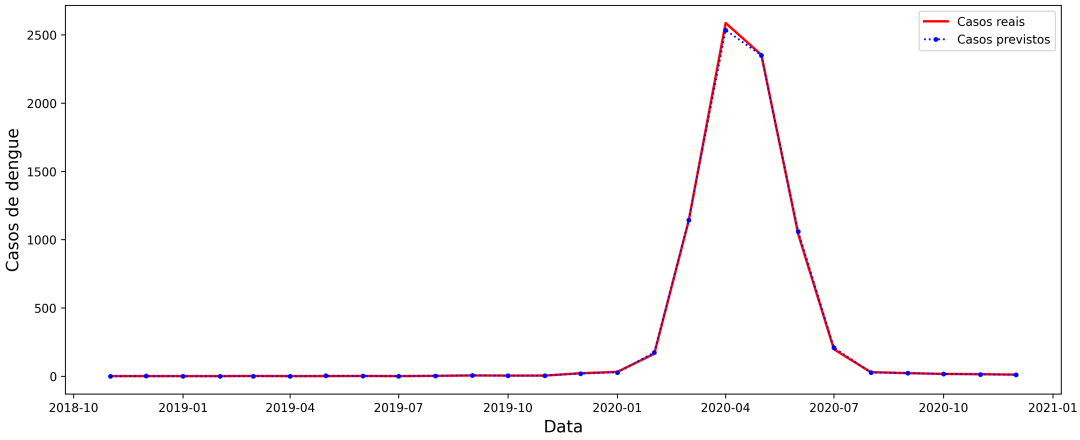
Ao consolidar os resultados alcançados, pode-se perceber que os modelos de aprendizagem de máquina apresentaram melhores performances em comparação aos modelos estatísticos. Isso se dá devido ao fato de serem regressores que podem ser escalados para mais de uma variável de influência, não somente aos dados referentes à própria série. A comparação está representada na

Figura 28 – Previsão de casos de dengue do modelo RF



Fonte: o autor

Figura 29 – Previsão de casos de dengue do modelo *Bagging*



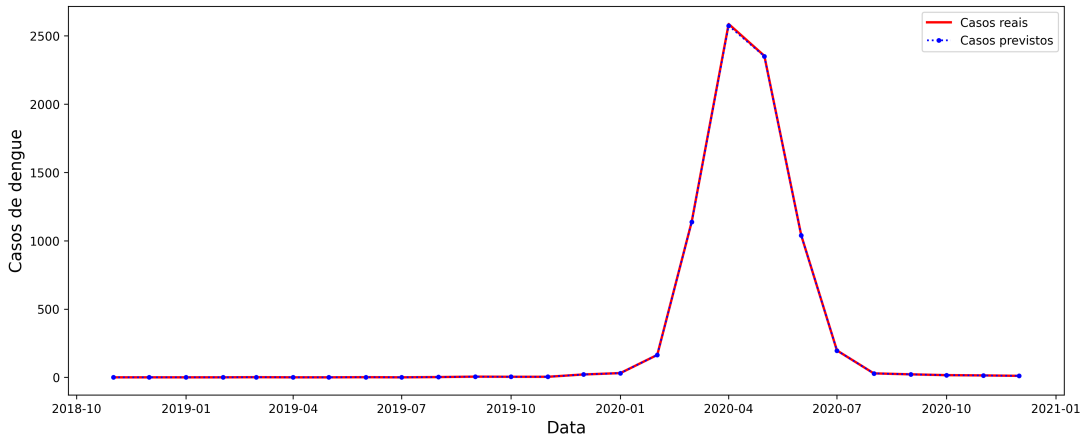
Fonte: o autor

Tabela 13 e, graficamente, na Figura 32. Na ilustração são representadas as curvas de estimação do melhor modelo estatístico (Média Móvel na cor azul com padrão +), do melhor modelo monolítico de aprendizagem de máquina (SVM em verde com padrão △) e, em amarelo com o padrão ∘, o melhor modelo baseado em conjuntos de classificadores (XGBoost Tree). Além disso, a linha vermelha contínua apresenta a curva do número de casos reais observados.

As curvas do SVM, do XGBoost Tree e dos casos reais estão praticamente sobrepostas, demonstrando a precisão alcançada por tais modelos. A diferença de desempenho e identificação do melhor método só foi possível analisando-se os valores de RMSE presentes na Tabela 13 onde percebe-se que o melhor desempenho global foi do SVM.

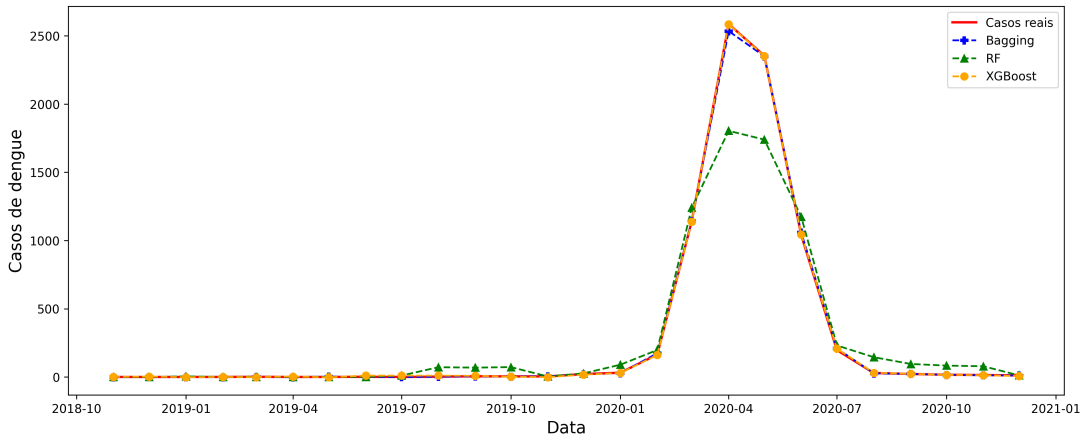
Visando ranquear os métodos implementados, na Figura 33 são apresentados os valores

Figura 30 – Previsão de casos de dengue do modelo XGBoost



Fonte: o autor

Figura 31 – Previsão de casos de dengue dos modelos de aprendizagem de máquina - técnicas de agrupamento



Fonte: o autor

de RMSE para todas as estratégias testadas neste estudo. Percebe-se que o SVM obteve o melhor desempenho geral, seguido pelas duas abordagens do XGBoost, primeiro o baseado em árvores e o segundo em classificadores lineares. O quarto melhor método foi o *bagging* combinado com perceptrons de múltiplas camadas e, em quinto lugar o MLP trabalhando de forma monolítica.

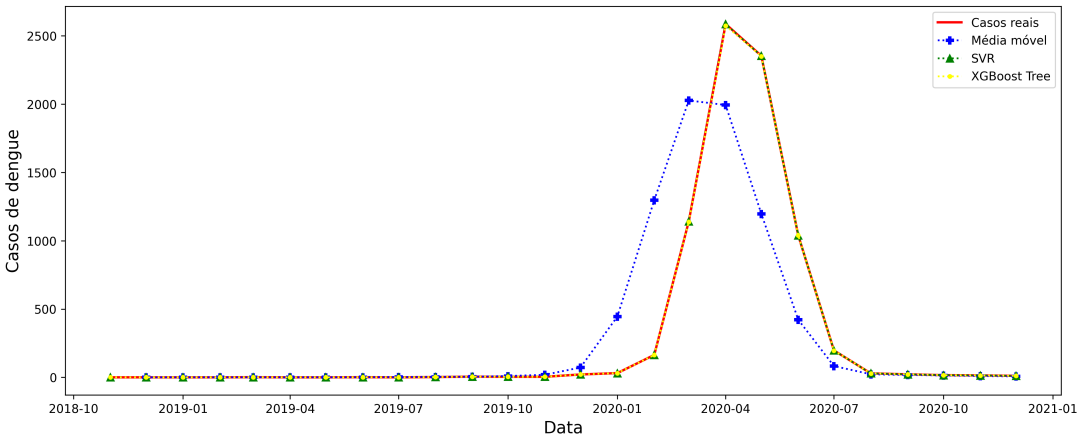
As três piores abordagens avaliadas foram o ARIMA (17º no ranking), seguido pelo *bagging* em conjunto com o SVM (18º pior desempenho) e, por fim, o pior desempenho global, a suavização exponencial.

Na Figura 34 é ilustrado o comportamento das cinco abordagens de melhor desempenho, conforme visto no ranking presente na Figura 33. Percebe-se a precisão das estratégias SVM (curva azul e padrão +), XGBoost nas suas duas configurações (denotadas pelas curvas amarela

Tabela 13 – Comparação do desempenho do melhor modelo estatístico, de aprendizagem de máquina monolítico e de técnicas de agrupamento de modelos

	Média Móvel	**SVM**	**XGBoost**
MAE	192,962	0,091	0,798
MSE	166.552,321	0,009	6,740
RMSE	408,108	0,093	2,596

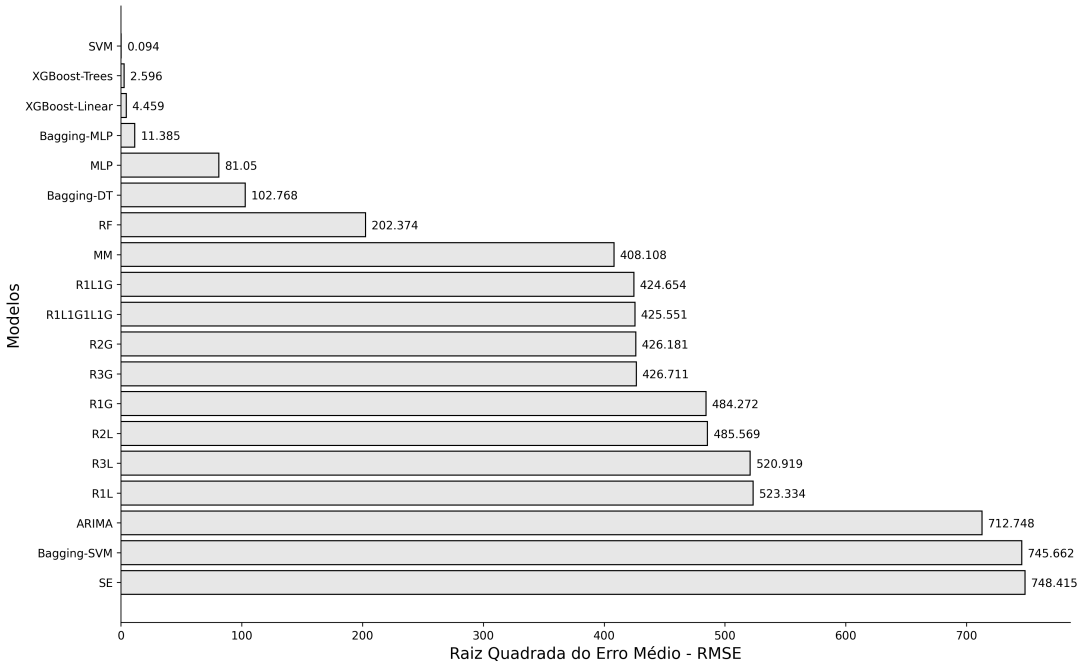
Figura 32 – Previsão de casos de dengue dos melhores modelos de cada estratégia



Fonte: o autor

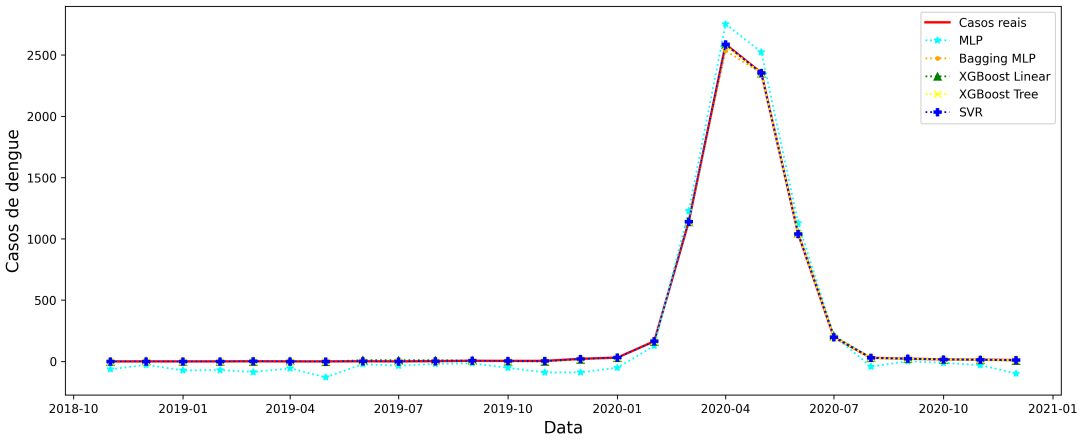
e verde, com os padrões $\times$ e $\triangle$, respectivamente). As curvas do modelo *Bagging* e do MLP podem ser percebidas mais facilmente pois não há tanta sobreposição, visto que suas métricas foram um pouco mais elevadas em comparação aos outros modelos deste rank. Os modelos estão representados pela curva laranja com padrão $\bullet$ e pela curva azul clara com padrão $\star$, respectivamente.

Figura 33 – Comparação do RMSE obtido por todos os modelos implementados



Fonte: o autor

Figura 34 – Comparação do desempenho das cinco melhores abordagens



Fonte: o autor

5

Considerações Finais

A dengue, que é transmitida pela picada do mosquito *Aedes aegypti*, é uma das mais importantes doenças dos países subdesenvolvidos e mostra-se uma grande preocupação urbana, principalmente na ocorrência de surtos da doença (RIBEIRO et al., 2006). Dessa forma, a previsão do número de casos da dengue possibilita a tomada de ações preventivas, para um controle efetivo da doença, tanto no combate ao vetor, mitigando o número de casos, quanto no tratamento dos infectados, além de preparar o sistema público de saúde para tal evento (GHARBI et al., 2011).

Neste contexto, no presente estudo fez-se o levantamento de tópicos interessantes, em relação à série temporal das ocorrências de dengue, e analisar modelos que pudessem ser eficientes na predição do número de casos de dengue, mais especificamente na cidade de Cascavel, Paraná.

Durante a análise da série temporal, foi verificada a presença de sazonalidade e tendências na série temporal de dados, bem como a correlação entre o número de casos de dengue e variáveis climáticas, uma vez que o mosquito transmissor se prolifera em acúmulos de água parada. Além disso, foi estudado um possível impacto em retrocesso das variáveis climáticas, para isso fez-se a correlação das variáveis deslocadas e em diferentes granularidades: diária, semanal e mensal.

Visando avaliar a aplicabilidade de modelos de estimação do número de casos de dengue, foram implementados e avaliados nove modelos de predição, sendo três deles abordagens estatísticas e o restante abordagens de aprendizagem de máquina. A seguir, serão apresentadas as principais conclusões decorrentes do estudo, apontando a eficiência dos modelos e também, elencando os pontos de relevância da série de casos de dengue, em Cascavel.

A série temporal mostrou possuir uma sazonalidade bastante forte entre os primeiros meses do ano, onde há registro da maioria dos casos confirmados em determinado período. Tal fato é ilustrado nas Figuras 10 e 11. Os meses de março e abril, em especial, registram cerca de 65% dos casos. Já os meses de fevereiro e maio também possuem um número significativo, registrando aproximadamente 27% do total de casos em um ano. Esse comportamento revela um

pico comum em praticamente todos os anos estudados, onde o número de casos começa a subir a partir de fevereiro, atingindo seu máximo entre março e abril e voltando a diminuir no mês de maio, assim como representado na Figura 12.

Corroborando com a representação visual dos dados, o teste de periodicidade de Rayleigh realizado rejeitou a hipótese nula que indica homogeneidade na distribuição dos dados e o valor obtido pelo teste indica que a série não possui comportamento regular, em um nível de 5% de significância. Tendo em vista que a série não é homogênea, pode-se concluir que ela não é estacionária e o cálculo do vetor circular médio P resultou em uma intensidade de $0,8089$, visto que este escalar é compreendido entre $[0, 1]$, indica que há uma direção específica e intensa da quantidade de casos registrados.

A partir da análise conjunta do teste de Rayleigh e do valor de P, pode-se afirmar que há sazonalidade na ocorrência dos casos da doença e que estes estão concentrados em uma determinada época do ano (entre fevereiro e maio).

Uma vez que as variáveis climáticas foram adotadas como suporte para a predição do número de casos de dengue (temperatura média, umidade relativa do ar e precipitação), é interessante analisar sua influência sobre a variável alvo. O estudo de correlação foi realizado com o objetivo de escolher uma configuração dos dados, que melhor se adapte ao comportamento visto na realidade.

A análise da correlação (Coeficiente de Correlação de Person) foi feita com granularidade diária, semanal e mensal. Os valores obtidos indicaram que a maior relação foi observada com a amostragem mensal com deslocamento de 30 dias. Acredita-se que tal valor é justificado pelo fato de que o *Aedes* leva em torno de 20 a 30 dias para eclodir de seu ovo, se desenvolver e picar uma pessoa contaminada, estando após alguns dias, apto a contaminar outra pessoa.

Mesmo que os coeficientes ainda sejam próximos de zero, a influência das variáveis climáticas é notável no resultado dos modelos de predição, pois os modelos estatísticos e a rede neural recorrente utilizam somente o número de casos para realizar suas predições. Tais modelos não obtiveram as melhores métricas ao tentar prever os casos. Apesar disso, a média móvel e a RNN mostram curvas bastante interessantes quando se trata da previsão de picos da doença, como ocorre nos primeiros meses do ano.

Tendo em vista os dois grandes picos (2016 e 2020) observados durante a série estudada, tais fatos considerados exceções dentro da linha temporal do número de casos registrados em Cascavel, contribuíram para a queda de performance dos modelos estatísticos. Analisando o modelo ARIMA, percebeu-se que o modelo foi capaz de prever o aumento de casos onde normalmente ocorre, sendo influenciado pelos anos anteriores. Contudo, não houve comportamento parecido na época que abrange o conjunto de teste, ou seja, a magnitude do número de casos de dengue foi expressiva demais no ano de 2020 e o modelo não previu tal acontecimento anormal.

Em contrapartida, o desempenho dos modelos de aprendizagem de máquina mostram

que a predição do número de casos de dengue se torna muito mais precisa quando utilizam-se variáveis climáticas de influência além da própria série temporal. Tal fato é comprovado quando metrifica-se a predição realizada pelo modelo SVM, no qual, obteve-se valores muito próximos de zero, para as três métricas adotadas. Além do SVM que mostrou um desempenho bastante próximo do real outras estratégias puderam apresentar desempenhos bastante precisos. Tais métodos foram as estratégias baseadas em conjuntos de regressos. O primeiro, *Bagging* trabalhando com grupo de MLPs e o segundo, *Extreme Gradient Boosting* empregando grupos de classificadores lineares e baseados em árvores de decisão.

As soluções encontradas se mostraram bastante positivas, tanto pela performance que apresentaram, quanto pela sua viabilidade, uma vez que não envolvem dados complexos e que conseguem expressar por si só a realidade do local onde o estudo foi aplicado.

Cabe destacar que, mesmo com a menor incidência de chuvas em Cascavel no ano de 2020, o que acarreta em uma menor umidade relativa do ar, o número de casos foi extremamente alto em comparação ao restante da série temporal abordada. O total de casos observados no período de janeiro de 2007 à dezembro de 2019, não atinge 35% dos casos registrados somente no ano de 2020. Além disso, neste ano foi decretado estado de pandemia por conta da COVID-19, tal fato voltou a maior atenção do governo e do sistema público de saúde para a nova doença em ascensão, o que contribuiu para um possível relaxamento dos cuidados básicos, por parte da população, e descontrole da dengue.

Diversas medidas de isolamento social foram tomadas para conter o avanço da COVID-19, dentre as quais, a visita dos agentes de endemias foi suspensa na cidade de Cascavel e os moradores não podiam receber os funcionários da vigilância (G1, 2022c). Acredita-se que muitos casos que antes eram subnotificados, foram registrados devido à similaridade de alguns sintomas entre a dengue e a COVID-19. Mesmo com o surgimento e preocupação por uma nova doença, mostra-se necessário manter as ações tomadas para evitar surtos de dengue, como ocorrido em 2020. Entretanto, tal fato navega em um dilema muito forte entre decidir qual o mal menor, neste período, devido à suspensão de atividades presenciais em toda a cidade.

A maioria dos modelos avaliados neste trabalho sofreram com as grandes variações que os anos de 2016 e 2020 apresentaram, pois o ano de 2020 estava presente no conjunto de teste, o que contribuiu para prejudicar as métricas de erro de modelos considerados robustos e frequentemente utilizados na literatura, como o ARIMA e as redes neurais recorrentes.

Após analisar os resultados obtidos pelos experimentos descritos no Capítulo 4, é possível concluir que a predição do número de casos de dengue pode ser realizada utilizando tanto a série temporal da doença, quanto as variáveis climáticas eleitas para esse estudo. Entretanto, os modelos que utilizaram as variáveis de apoio mostraram uma performance mais precisa que os baseados somente na série temporal. Mesmo com o pico anormal observado no ano de 2020, os modelos SVM, XGBoost, *Bagging* e MLP foram capazes de prevê-lo.

Para trabalhos futuros, ainda é possível adicionar mais variáveis ambientais de apoio como o LIRAa, índice de vegetação (PHAM et al., 2015), temperatura da superfície do mar (quando for cabível) (LAUREANO-ROSARIO et al., 2018), ponto de orvalho e incidência solar (ZHU; HUNTER; JIANG, 2016). Além disso, outras métricas sociais como a população (LAUREANO-ROSARIO et al., 2018), sexo, raça e profissão como visto no trabalho de Tarmizi et al. (2013) podem ser empregadas no processo de análise e estimação.

Séries temporais são estendidas com o passar do tempo. Este trabalho utilizou o período de 2007 a 2020 como foco de estudo, sendo assim, sempre será possível estender a base de dados ao longo dos anos, adicionando novos dados aos modelos já apresentados ou mesmo incluir novas estratégias de estimação, com o objetivo de disseminar a importância da predição de casos de dengue em Cascavel e no Brasil como um todo. Além disso, pode-se avaliar a aplicabilidade dos modelos construídos sobre outras áreas de estudo, com características semelhantes a Cascavel bem como em áreas com configuração climática distinta da cidade.

Referências

ABURAS, H. M.; CETINER, B. G.; SARI, M. Dengue confirmed-cases prediction: A neural network model. *Expert Systems with Applications*, v. 37, p. 4257–4260, Abril 2010. Disponível em <https://doi.org/10.1016/j.eswa.2009.11.077>. Citado na página 35.

AHMAD, R. et al. Factors determining dengue outbreak in malaysia. *PLoS ONE*, v. 13, n. 2, Fevereiro 2018. Disponível em <https://doi.org/10.1371/journal.pone.0193326>. Citado na página 32.

AZHAR, K.; MARINA, R.; ANWAR, A. A prediction model of dengue incidence using climate variability in denpasar city. *Health Science Journal of Indonesia*, v. 8, n. 2, p. 68–73, Dezembro 2017. Disponível em <http://dx.doi.org/10.22435/hsji.v8i2.6952.68-73>. Citado 2 vezes nas páginas 31 e 36.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: *Computation and Language*. [s.n.], 2014. Disponível em: <https://arxiv.org/abs/1409.0473>. Citado na página 26.

BAQUERO, O. S.; SANTANA, L. M. R.; NETO, F. C. Dengue forecasting in são paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PLoS ONE*, v. 13, n. 4, Abril 2018. Disponível em <https://doi.org/10.1371/journal.pone.0195065>. Citado 2 vezes nas páginas 15 e 31.

BENGIO, Y.; BOULANGER-LEWANDOWSKI, N.; PASCANU, R. Advances in optimizing recurrent networks. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2013. p. 8624–8628. Disponível em <https://doi.org/10.1109/ICASSP.2013.6639349>. Citado na página 26.

Biblioteca Virtual em Saúde do Ministério da Saúde. *Dicas em Saúde - Dengue*. 2007. Disponível em: <https://bvsms.saude.gov.br/bvs/dicas/33dengue.html>. Acesso em: 12 jul 2022. Citado na página 30.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 1. ed. [S.l.]: Springer, 2006. Citado na página 23.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis Forecasting and Control*. 5. ed. New Jersey: Jon Wiley & Sons, 2016. Citado na página 18.

BRAGA, O. C. et al. A mobile health solution for diseases control transmitted by aedes aegypti mosquito using predictive classifiers. In: *CoUrb*. Porto Alegre: [s.n.], 2017. p. 144–156. Disponível em <https://sol.sbc.org.br/index.php/courb/article/view/2570>. Citado 2 vezes nas páginas 15 e 22.

BRAZIER, K. T. S. Confidence intervals from the Rayleigh test. *Monthly Notices of the Royal Astronomical Society*, v. 268, n. 3, p. 709–712, 06 1994. ISSN 0035-8711. <https://doi.org/10.1093/mnras/268.3.709>. Citado na página 43.

BREIMAN, L. Bagging predictors. *Mach Learn*, v. 24, p. 123–140, 08 1996. <https://doi.org/10.1007/BF00058655>. Citado 2 vezes nas páginas 28 e 29.

BREIMAN, L. et al. *Classification and Regression Trees*. 1. ed. [S.l.]: Chapman and Hall/CRC, 1984. Citado na página 24.

BROWNLEE, J. *Deep Learning for Time Series Forecasting*. 1.4. ed. [S.l.]: Machine Learning Mastery, 2018. Citado 2 vezes nas páginas 22 e 26.

CARLOS, M. A.; NOGUEIRA, M.; MACHADO, R. J. Analysis of dengue outbreaks using big data analytics and social networks. In: *The 2017 4th International Conference on Systems and Informatics*. [S.l.: s.n.], 2017. p. 1592–1597. Disponível em <https://doi.org/10.1109/ICSAI.2017.8248538>. Citado na página 32.

CBNLONDRINA. *Paraná registra alta expressiva nos casos de dengue no atual período de monitoramento da doença*. 2022. CBN Londrina 100,9FM. Disponível em <http://tinyurl.com/2p9y4hk4>. Acesso em: 12 jul 2022. Citado na página 14.

CGN. *Cascavel chega aos 7.145 casos de dengue*. 2022. Central Gazeta de Notícias. Disponível em <http://tinyurl.com/2vua9hth>. Acesso em: 12 jul 2022. Citado na página 14.

CHATFIELD, C. *Time-Series Forecasting*. 1. ed. Boca Raton, FL, USA: Chapman & Hall, 2000. Citado 2 vezes nas páginas 18 e 19.

CHATFIELD, C. *The Analysis of Time Series - An Introduction*. 6. ed. [S.l.]: Chapman and Hall/CRC, 2003. Citado 2 vezes nas páginas 17 e 18.

CHUNG, J. et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: *Neural and Evolutionary Computing*. [S.l.: s.n.], 2014. Disponível em <https://arxiv.org/abs/1412.3555>. Citado na página 26.

Data Science Team. *Gradient Boosting – O que você precisa de saber*. 2020. Disponível em: <https://datascience.eu/pt/aprendizado-de-maquina/gradient-boosting-o-que-voce-precisa-de-saber/>. Acesso em: 15 ago 2022. Citado na página 29.

FERNANDO, J. *Moving Average*. 2021. Disponível em: <https://www.investopedia.com/terms/m/movingaverage.asp>. Acesso em: 12 jul 2022. Citado na página 19.

Fiocruz. *Glossário de doença - Dengue*. 2013. Disponível em: <https://agencia.fiocruz.br/dengue-0>. Acesso em: 12 jul 2022. Citado na página 30.

FRIEDMAN, J. H. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, v. 38, n. 4, p. 367–378, 2002. ISSN 0167-9473. Disponível em <https://doi.org/10.1016/S0167-9473(01)00065-2>. Citado na página 29.

G1. *Cascavel decreta epidemia de dengue após mais de mil casos confirmados no atual ano epidemiológico*. 2022. Oeste e Sudoeste RPC. Disponível em <http://tinyurl.com/468b5kw5>. Acesso em: 12 jul 2022. Citado na página 14.

G1. *Cascavel tem alto risco de infestação de dengue, aponta levantamento*. 2022. Oeste e Sudoeste RPC. Disponível em <https://tinyurl.com/y79cxv99>. Acesso em: 12 jul 2022. Citado na página 31.

G1. *Moradores de Cascavel com Covid-19 ou sintomas não devem receber agentes de endemias*. 2022. Pararná RPC - Vídeos. Disponível em <https://tinyurl.com/yc5vmsvh>. Acesso em: 12 jul 2022. Citado na página 63.

G1PR. *Prefeitura de Cascavel realiza ação de combate à dengue em bairros com maior índice de infestação*. 2021. G1 Oeste e Sudoeste RPC. Disponível em: <http://tinyurl.com/468b5kw5>. Acesso em: 12 jul 2022. Citado na página 31.

GALDI, P.; TAGLIAFERRI, R. Data mining: Accuracy and error measures for classification and prediction. In: ____. [S.l.: s.n.], 2018. ISBN 9780128096338. Citado na página 28.

GHARBI, M. et al. Time series analysis of dengue incidence in guadeloupe, french west indies: Forecasting models using climate variables as predictors. *BMC infectious diseases*, v. 11, p. 166, Junho 2011. Disponível em <https://doi.org/10.1186%2F1471-2334-11-166>. Citado 2 vezes nas páginas 15 e 61.

GOODFELLOW, I.; YOSHUA, B.; COURVILLE, A. *Deep Learning (Adaptive Computation and Machine Learning series)*. Illustrated. [S.l.]: The MIT Press, 2016. Citado na página 26.

GUNES, V. et al. Combination, cooperation and selection of classifiers: A state of the art. *International Journal of Pattern Recognition and Artificial Intelligence*, v. 17, n. 08, p. 1303–1324, 2003. Disponível em <https://doi.org/10.1142/S0218001403002897>. Citado na página 27.

GUO, P. et al. Developing a dengue forecast model using machine learning: A case study in china. *PLoS Neglected Tropical Diseases*, v. 11, n. 10, Outubro 2017. Disponível em <https://doi.org/10.1371/journal.pntd.0005973>. Citado 3 vezes nas páginas 22, 31 e 35.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 11 1997. ISSN 0899-7667. Disponível em <https://doi.org/10.1162/neco.1997.9.8.1735>. Citado na página 26.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 2. ed. Monash University, Australia: Texts, 2018. Citado na página 21.

IBGE. *Brasil/Paraná/Cascavel - Panorama*. 2021. Disponível em: <https://cidades.ibge.gov.br/brasil/pr/cascavel/panorama>. Acesso em: 12 jul 2022. Citado na página 30.

KERAS. *Redes Neurais Recorrentes (RNN) com Keras*. 2022. Disponível em: <https://www.tensorflow.org/guide/keras/rnn>. Acesso em: 12 jul 2022. Citado na página 25.

KITTLER, J. Combining classifiers: A theoretical framework. *Pattern Analysis and Applications*, v. 1, n. 1, p. 18–27, 1998. ISSN 14337541. Disponível em <https://link.springer.com/article/10.1007/BF01238023>. Citado na página 27.

LASERNA, A. et al. Economic impact of dengue fever in latin america and the caribbean: a systematic review. *Revista Panamericana de Salud Pública*, v. 42, n. 17, p. 1–9, Setembro 2018. Disponível em <https://doi.org/10.26633/RPSP.2018.111>. Citado na página 14.

LAUREANO-ROSARIO, A. E. et al. Application of artificial neural networks for dengue fever outbreak predictions in the northwest coast of yucatan, mexico and san juan, puerto rico. *Tropical Medicine and Infectious Disease*, v. 3, n. 1, p. 1–16, Janeiro 2018. Disponível em <https://doi.org/10.3390/tropicalmed3010005>. Citado 3 vezes nas páginas 31, 32 e 64.

LE, J. *R Decision Tree Tutorial*. 2018. Disponível em: <https://www.datacamp.com/tutorial/decision-trees-R>. Acesso em: 12 jul 2022. Citado na página 25.

LEE, C.-H.; YANG, H.-C.; LIN, S.-J. Incorporating big data and social sensors in a novel early warning system of dengue outbreaks. In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Paris, França: [s.n.], 2015. p. 1428–1433. Disponível em <https://doi.org/10.1145/2808797.2808883>. Citado na página 32.

MAKRIDAKIS, S. G.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. *Forecasting: Methods and Applications*. 3. ed. Nova Iorque, USA: Wiley, 1997. Citado 3 vezes nas páginas 19, 20 e 21.

MILLS, T. C. *Inteligência Artificial*. 1. ed. Londres: Elsevier, 2019. Citado na página 18.

MITTELMANN, M.; SOARES, D. G. Previsão de casos de dengue em itajaí - sc por meio de redes neurais artificiais multicamadas e recorrentes. In: . [S.l.: s.n.], 2017. Citado 2 vezes nas páginas 31 e 35.

MITTELMANN, M.; SOARES, D. G. Previsão de casos de dengue no município de Guarulhos com redes neurais artificiais multicamadas e recorrentes. *Revista de Informática Aplicada*, v. 13, n. 2, p. 68–74, Fevereiro 2017. Disponível em <http://tinyurl.com/5n89hed8>. Citado 3 vezes nas páginas 15, 31 e 35.

MORADZADEH, A. et al. Performance evaluation of two machine learning techniques in heating and cooling loads forecasting of residential buildings. *Applied Sciences*, v. 10, n. 11, p. 1–12, 05 2020. Disponível em <https://doi.org/10.3390/app10113829>. Citado na página 24.

MORETTIN, P. A.; TOLOI, C. M. C. *Análise de Séries Temporais*. 2. ed. São Paulo: Blucher, 2006. Citado 3 vezes nas páginas 18, 19 e 20.

MUHILTHINI, P. et al. Dengue possibility forecasting model using machine learning algorithms. *International Research Journal of Engineering and Technology (IRJET)*, v. 05, n. 03, p. 1661–1665, 2018. ISSN 2395-0056. Disponível em: <https://www.irjet.net/archives/V5/i3/IRJET-V5I3373.pdf>. Citado na página 22.

NIELSEN, A. *Análise de Séries Temporais*. 1. ed. Rio de Janeiro: Alta Books, 2021. Citado na página 17.

PAHO. *Dengue y Dengue Grave - Casos y Muertes*. 2022. Disponível em: <http://tinyurl.com/yckurfdv>. Acesso em: 12 jul 2022. Citado na página 14.

PAHO. *Tasa de Incidencia por Dengue*. 2022. Disponível em: <http://tinyurl.com/52hck3rz>. Acesso em: 12 jul 2022. Citado na página 14.

PAHO. *Tasa de Letalidad por Dengue*. 2022. Disponível em: <http://tinyurl.com/yu2yrawn>. Acesso em: 12 jul 2022. Citado na página 14.

PHAM, D. N. et al. An efficient method to predict dengue outbreaks in kuala lumpur. In: *3rd International Conference on Artificial Intelligence and Computer Science*. Penang, Malaysia: [s.n.], 2015. p. 169–178. Disponível em <http://tinyurl.com/yc5ak4xf>. Citado 3 vezes nas páginas 32, 36 e 64.

PHAM, D. N. et al. A literature review of methods for dengue outbreak prediction. In: *The Eighth International Conference on Information, Process, and Knowledge Management*. Veneza, Itália: [s.n.], 2016. p. 7–13. Disponível em <http://tinyurl.com/mry442p6>. Citado 2 vezes nas páginas 15 e 22.

PHUNG, D. et al. Identification of the prediction model for dengue incidence in can tho city, a mekong delta area in vietnam. *Acta Tropica*, v. 141, p. 88–96, Outubro 2015. Disponível em <https://doi.org/10.1016/j.actatropica.2014.10.005>. Citado 2 vezes nas páginas 14 e 31.

Prefeitura Municipal de Cascavel. *Boletim da dengue: Cascavel registra 6.681 casos da doença*. 2020. Prefeitura Municipal de Cascavel - Notícias. Disponível em <http://tinyurl.com/mr2n73ae>. Acesso em: 12 jul 2022. Citado na página 31.

Prefeitura Municipal de Cascavel. *Clima - Mapa do Paraná*. 2021. Disponível em: <https://cascavel.atende.net/cidadao/pagina/mapas>. Acesso em: 12 jul 2022. Citado na página 30.

RIBEIRO, A. F. et al. Associação entre incidência de dengue e variáveis climáticas. *Revista de Saúde Pública*, v. 40, n. 4, p. 671–676, Agosto 2006. Disponível em <https://doi.org/10.1590/S0034-89102006000500017>. Citado 2 vezes nas páginas 14 e 61.

RIZZI, C. B. et al. Considerações sobre a dengue e variáveis de importância à infestação por aedes aegypti. *Revista Brasileira de Geografia Médica e da Saúde*, v. 13, n. 24, p. 24–40, Junho 2017. Disponível em <https://seer.ufu.br/index.php/hygeia/article/view/35133>. Citado na página 14.

Scikit-learn developers. *Decision Trees*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/tree.html#>. Acesso em: 12 jul 2022. Citado na página 24.

Scikit-learn developers. *Ensemble Methods*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/ensemble.html#forest>. Acesso em: 12 jul 2022. Citado na página 27.

Scikit-learn developers. *Neural Network Models (supervised)*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/neural_networks_supervised.html>. Acesso em: 12 jul 2022. Citado 2 vezes nas páginas 22 e 23.

Scikit-learn developers. *sklearn.ensemble.RandomForestRegressor*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>. Acesso em: 12 jul 2022. Citado na página 27.

Scikit-learn developers. *Support Vector Machines*. 2021. Disponível em: <https://scikit-learn.org/stable/modules/svm.html>. Acesso em: 12 jul 2022. Citado 2 vezes nas páginas 23 e 24.

SIMEPAR. *Sistema de Tecnologia e Monitoramento Ambiental do Paraná*. 2022. Sistema de Tecnologia e Monitoramento Ambiental do Paraná. Disponível em: <http://www.simepar.br/>. Acesso em: 12 jul 2022. Citado na página 35.

SMOLA, A. J.; SCHOLKOPF, B. A tutorial on support vector regression. In: *Statistics and Computing*. Holanda: [s.n.], 2004. p. 199–222. Disponível em <https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88>. Citado na página 23.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Datamining Mineração de Dados*. 1. ed. Rio de Janeiro, Brasil: Editora Ciência Moderna, 2009. Citado 5 vezes nas páginas 22, 23, 27, 28 e 29.

TARMIZI, N. D. A. et al. Classification of dengue outbreak using data mining models. *Research Notes in Information Science (RNIS)*, v. 12, n. 12, p. 71–75, 04 2013. ISSN 2287-1934. Disponível em <https://tinyurl.com/db6bjfaj>. Citado na página 64.

VEEN, F. V. *The Neural Network Zoo*. 2016. Disponível em: <https://www.asimovinstitute.org/neural-network-zoo/>. Acesso em: 12 jul 2022. Citado 2 vezes nas páginas 25 e 26.

WHO. *Dengue and severe dengue*. 2022. Disponível em: <https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue>. Acesso em: 12 jul 2022. Citado 2 vezes nas páginas 14 e 30.

WIKIPEDIA. *Cascavel (Paraná)*. 2022. Wikipedia. Disponível em: <https://pt.wikipedia.org/wiki/Cascavel_(ParanÃą)>. Acesso em: 12 jul 2022. Citado na página 31.

ZHOU, Z.-H. *Ensemble Methods - Foundations and Algorithms*. 1. ed. [S.l.]: Chapman and Hall/CRC, 2012. Citado 2 vezes nas páginas 27 e 29.

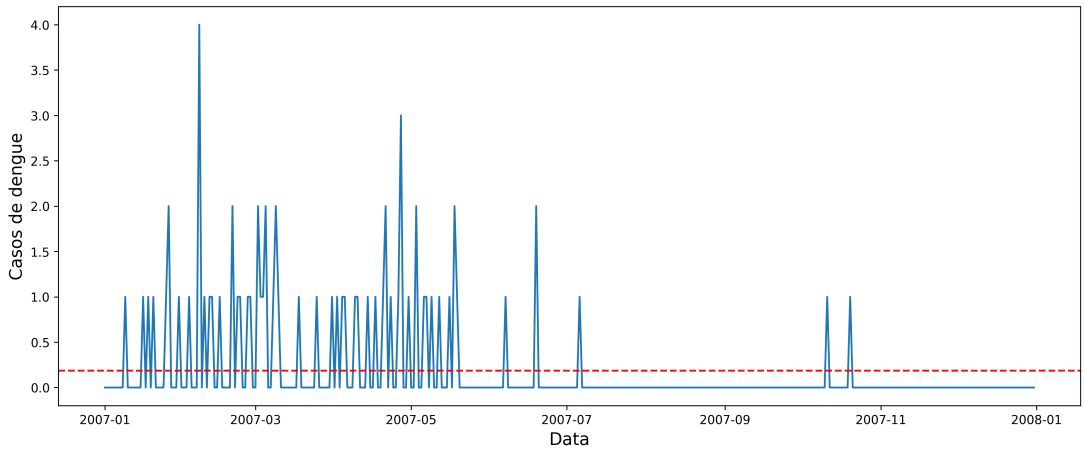
ZHU, G.; HUNTER, J.; JIANG, Y. Improved prediction of dengue outbreak using the delay permutation entropy. In: *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. Chengdu, China: [s.n.], 2016. p. 828–832. Disponível em <https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2016.172>. Citado 2 vezes nas páginas 32 e 64.

Apêndices

APÊNDICE A – Gráficos da quantidade de casos de dengue ao longo dos anos do período estudado

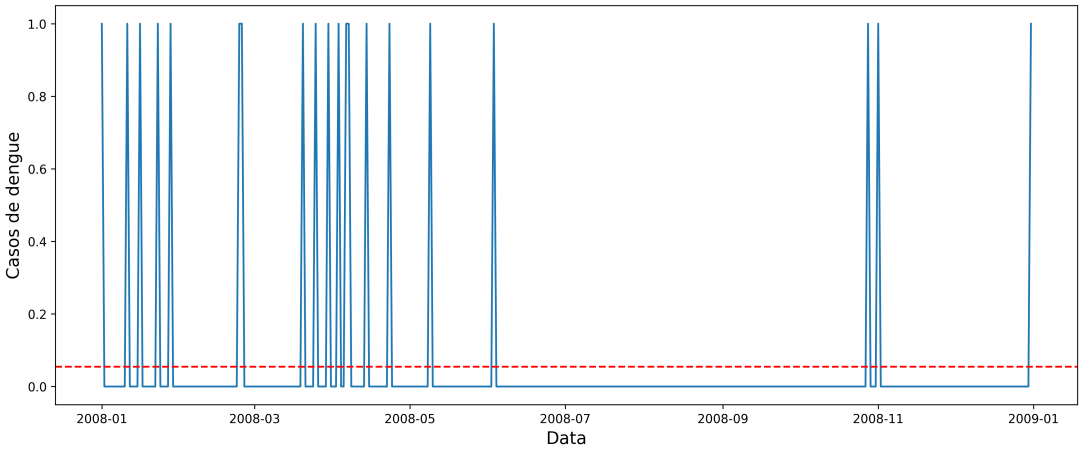
Os gráficos ilustrados nas Figuras 35 à 48 apresentam a curva com a distribuição do número de casos de dengue observados ao longo de cada ano especificado. Cada ponto representado no eixo das abscissas corresponde a um dia do ano. A reta em vermelho corresponde ao número de casos médio diários para o ano em análise.

Figura 35 – Quantidade de casos de dengue ao longo de 2007



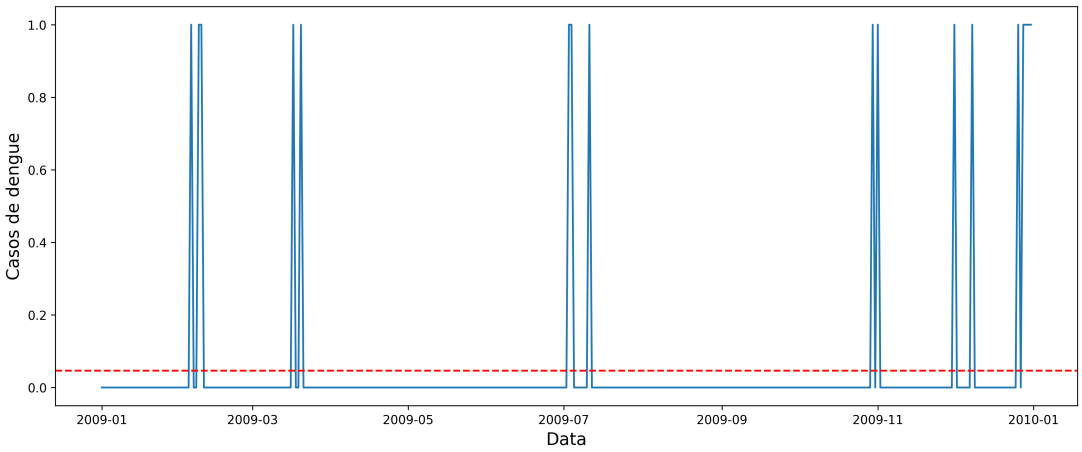
Fonte: o autor

Figura 36 – Quantidade de casos de dengue ao longo de 2008



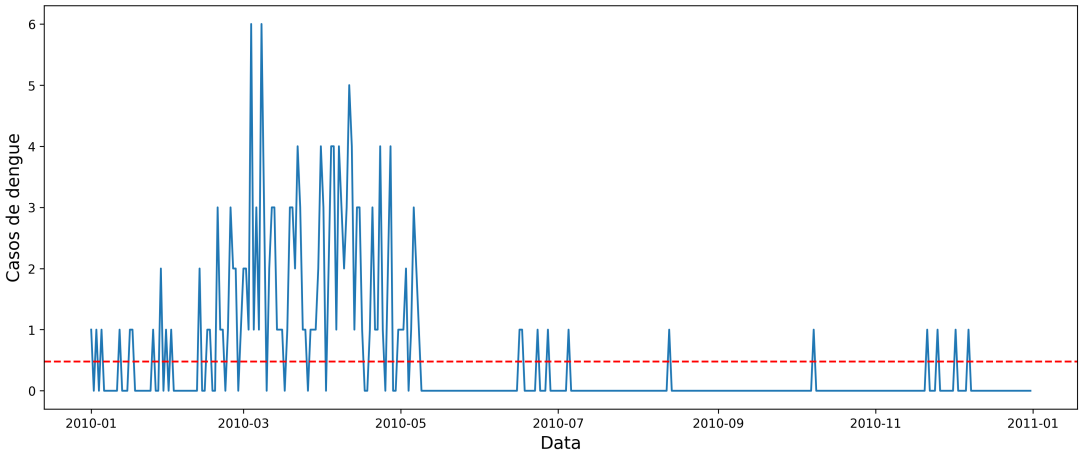
Fonte: o autor

Figura 37 – Quantidade de casos de dengue ao longo de 2009



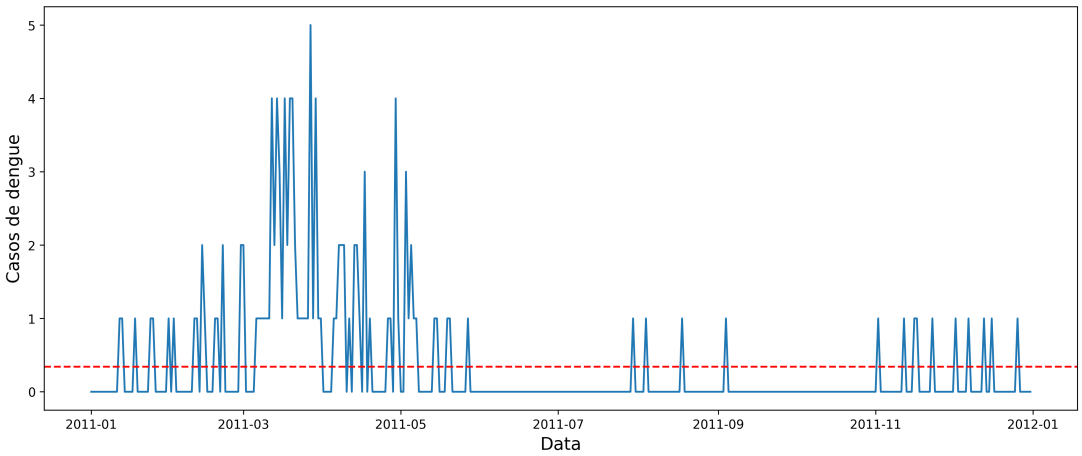
Fonte: o autor

Figura 38 – Quantidade de casos de dengue ao longo de 2010



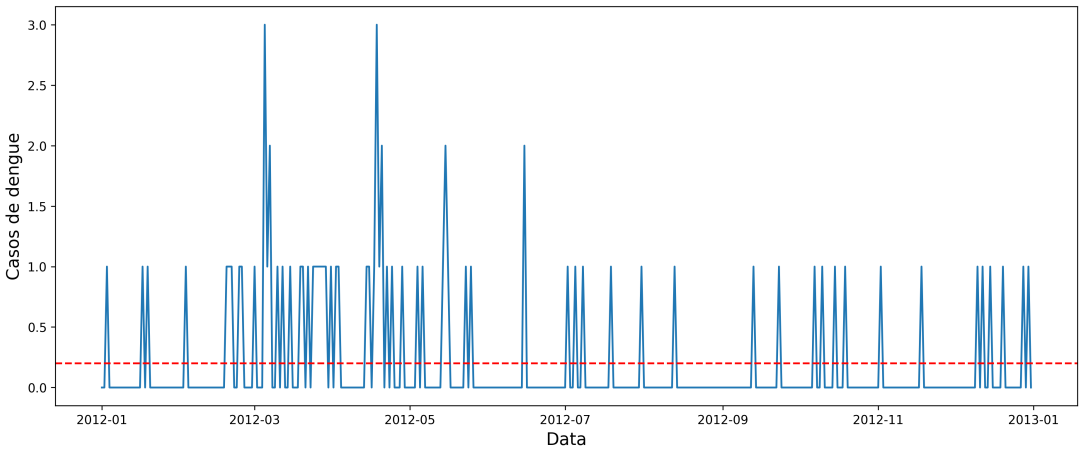
Fonte: o autor

Figura 39 – Quantidade de casos de dengue ao longo de 2011



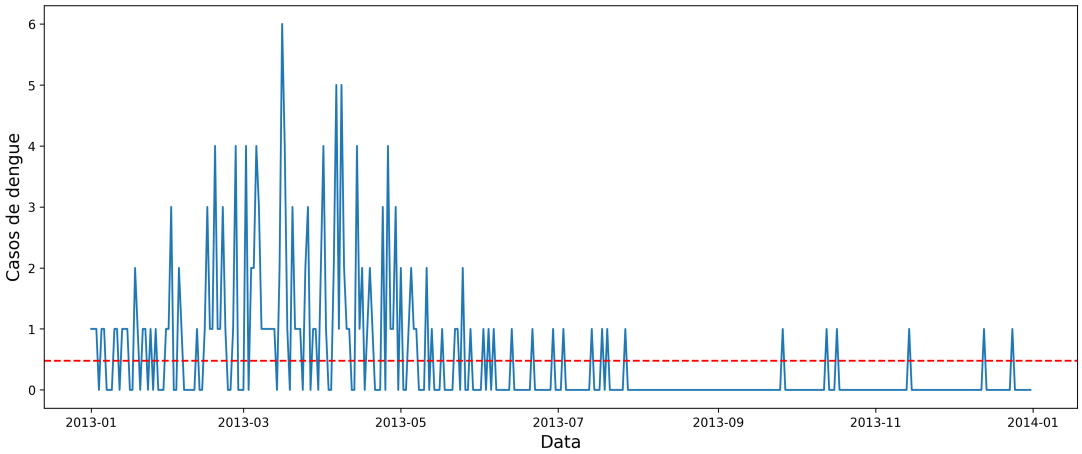
Fonte: o autor

Figura 40 – Quantidade de casos de dengue ao longo de 2012



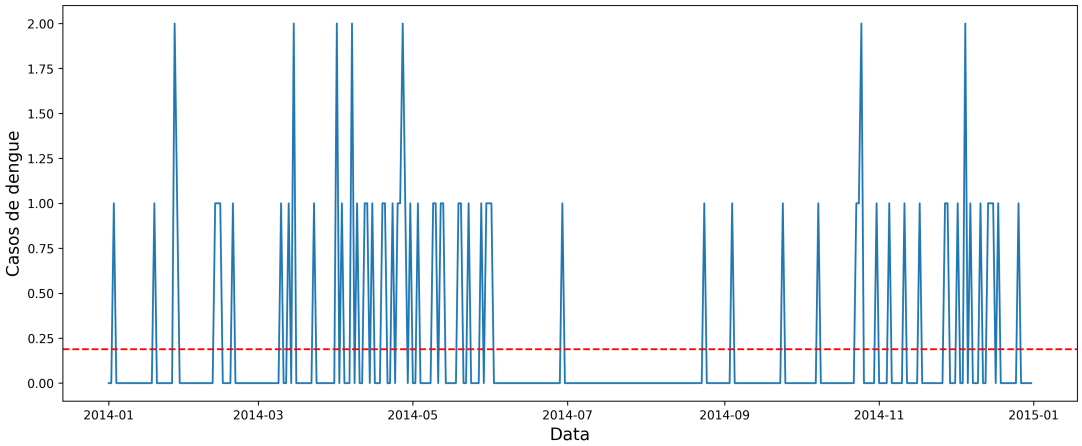
Fonte: o autor

Figura 41 – Quantidade de casos de dengue ao longo de 2013



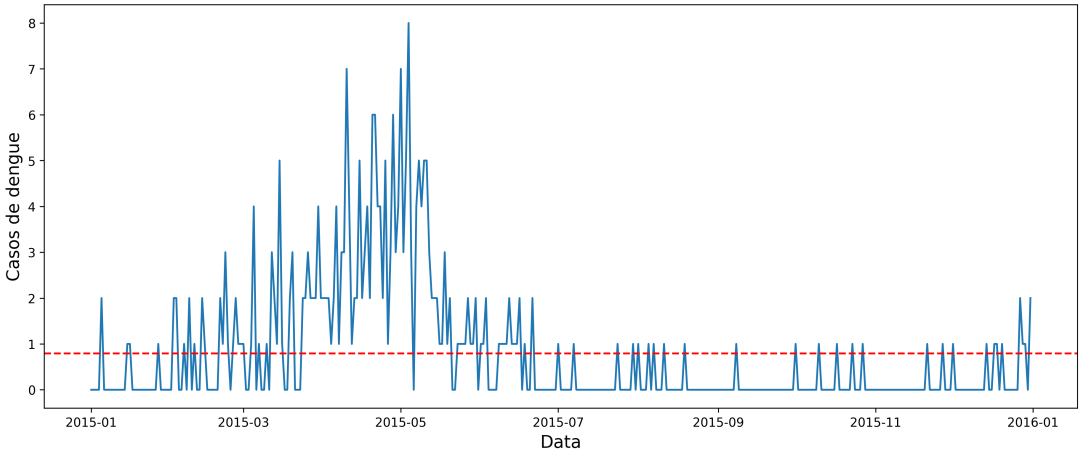
Fonte: o autor

Figura 42 – Quantidade de casos de dengue ao longo de 2014



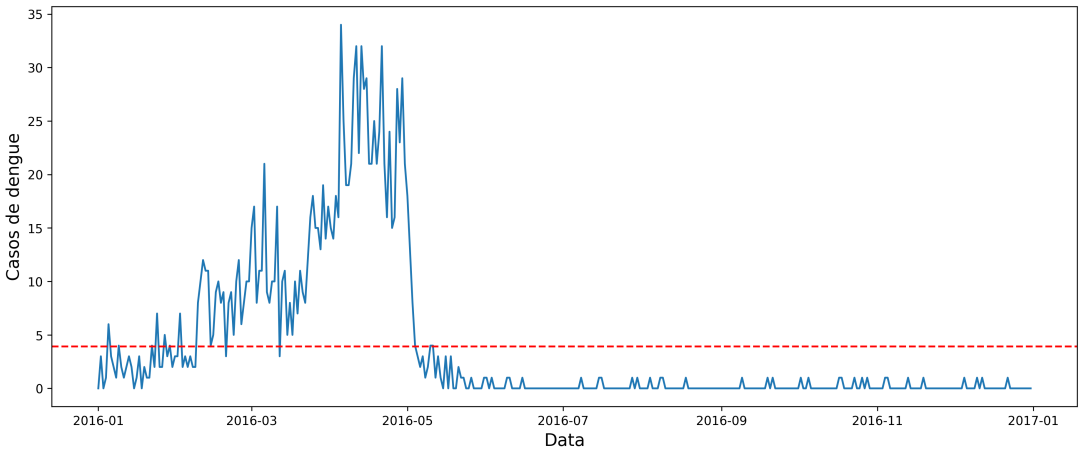
Fonte: o autor

Figura 43 – Quantidade de casos de dengue ao longo de 2015



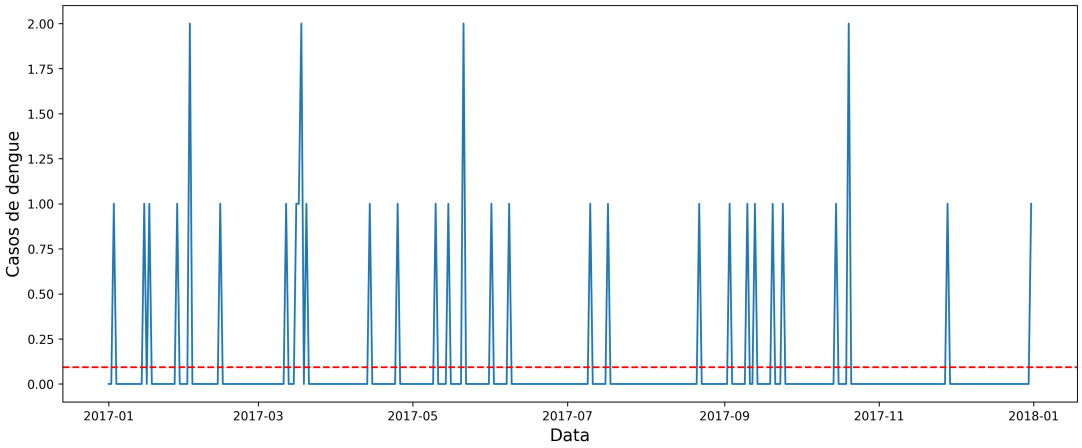
Fonte: o autor

Figura 44 – Quantidade de casos de dengue ao longo de 2016



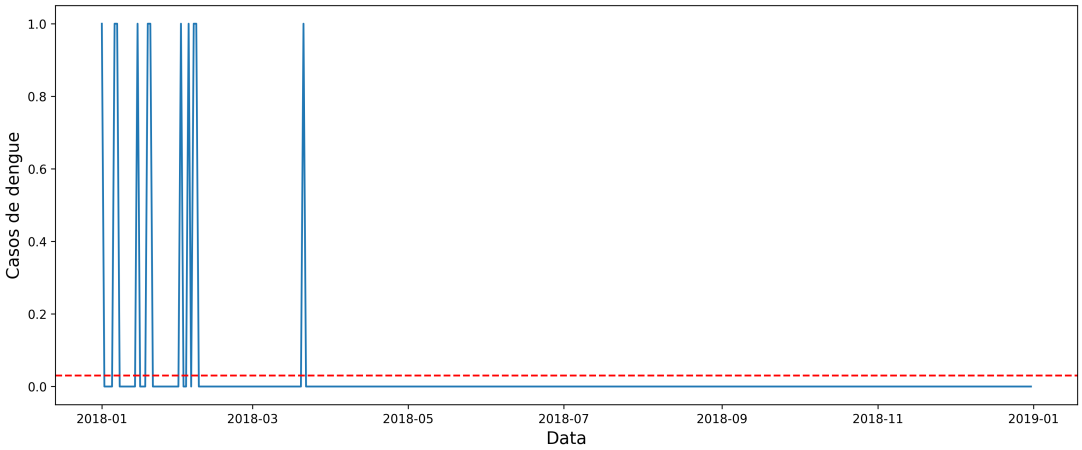
Fonte: o autor

Figura 45 – Quantidade de casos de dengue ao longo de 2017



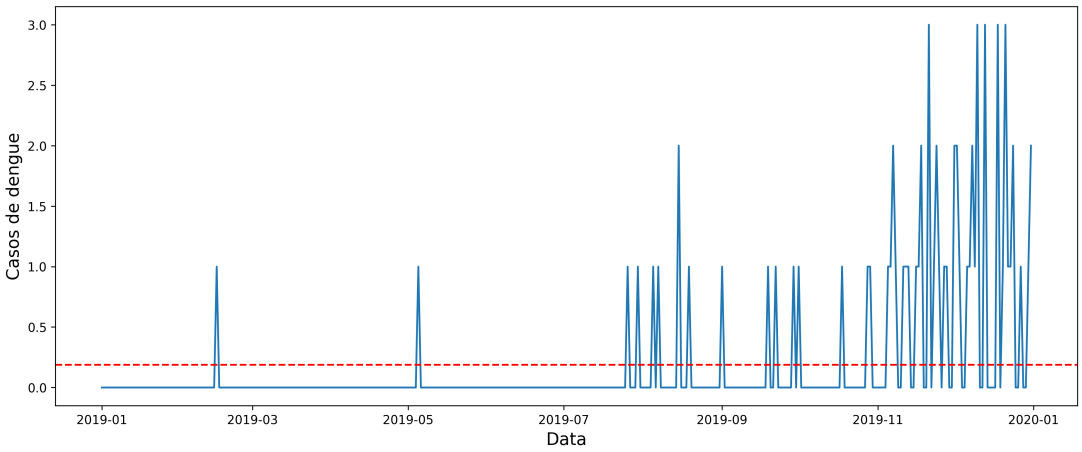
Fonte: o autor

Figura 46 – Quantidade de casos de dengue ao longo de 2018



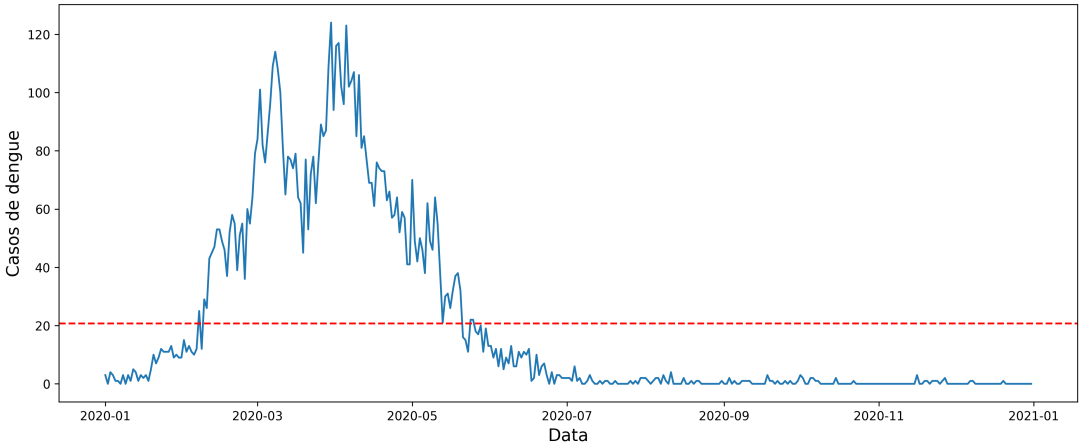
Fonte: o autor

Figura 47 – Quantidade de casos de dengue ao longo de 2019



Fonte: o autor

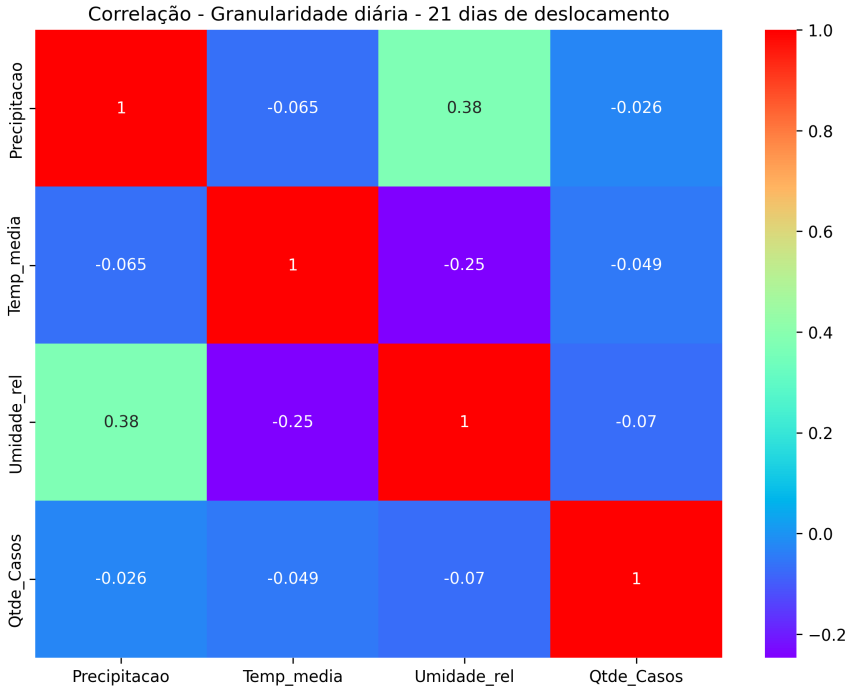
Figura 48 – Quantidade de casos de dengue ao longo de 2020



Fonte: o autor

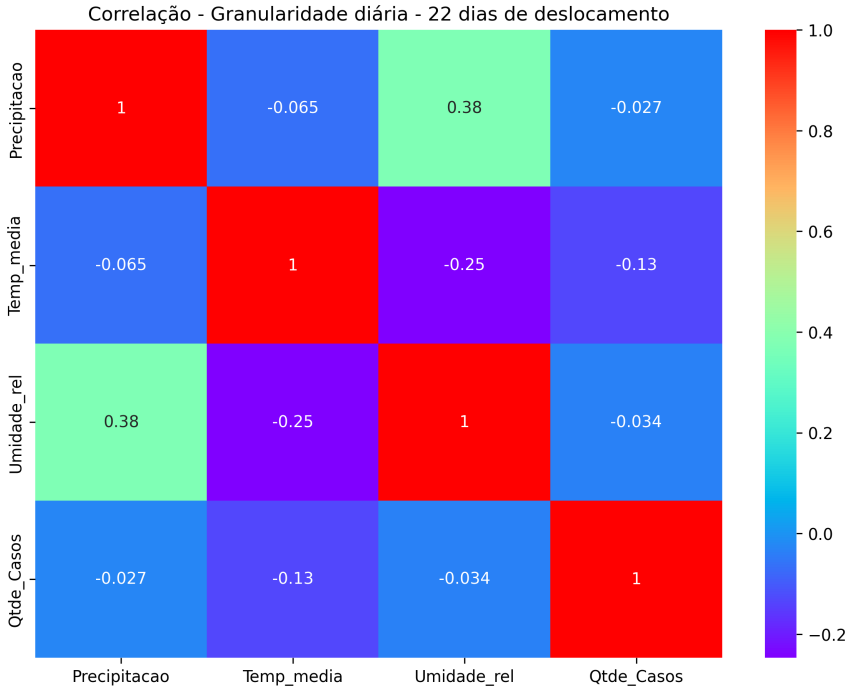
APÊNDICE B – Matrizes de confusão para correlação entre as variáveis climáticas e a quantidade de casos de dengue, com deslocamento em dias

Figura 49 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 21 dias



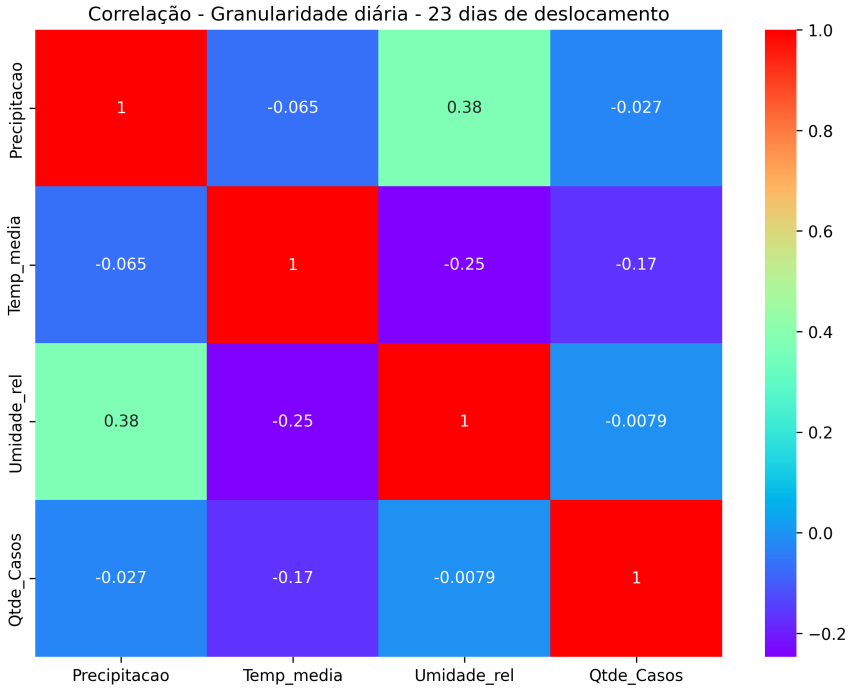
Fonte: o autor

Figura 50 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 22 dias



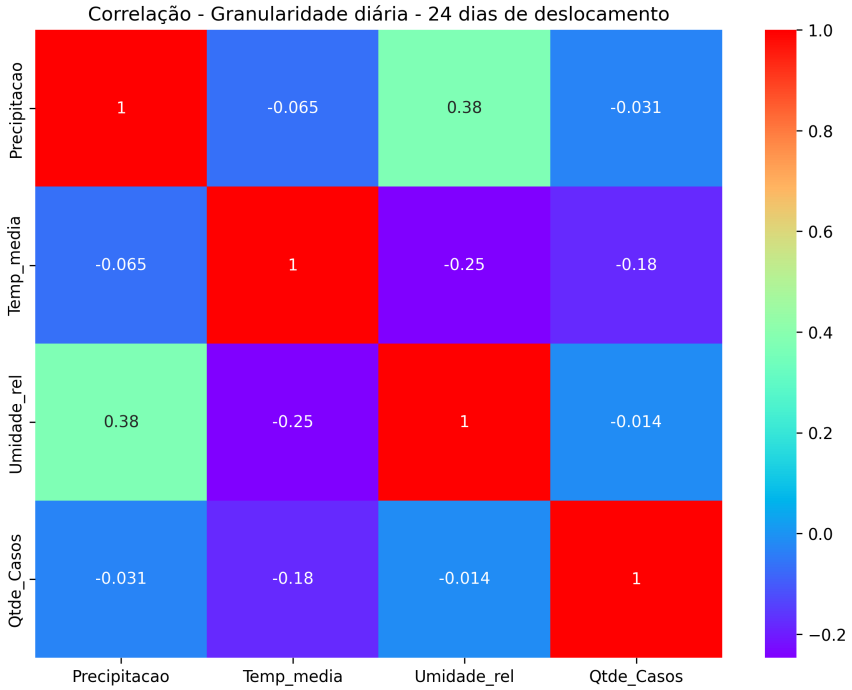
Fonte: o autor

Figura 51 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 23 dias



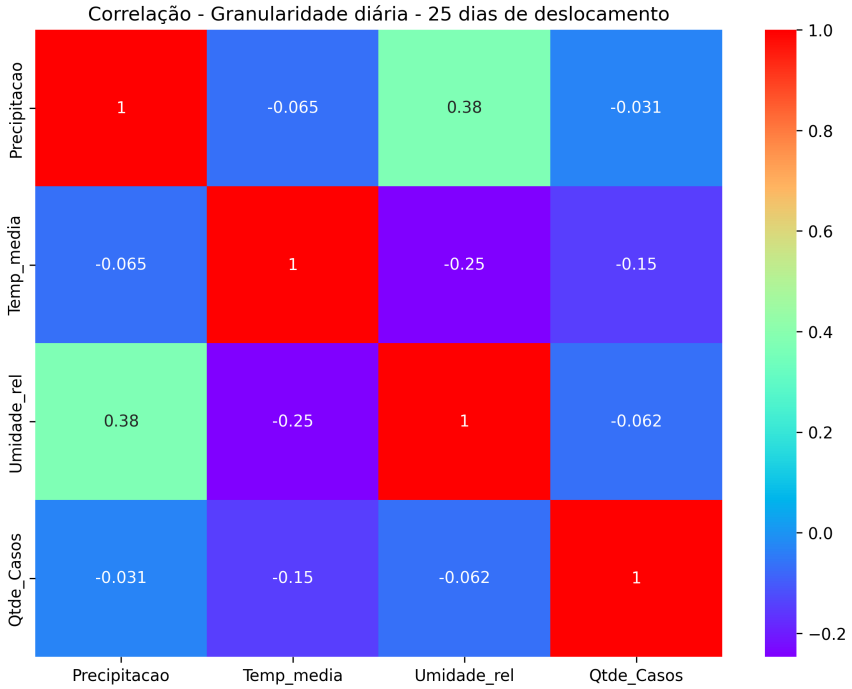
Fonte: o autor

Figura 52 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 24 dias



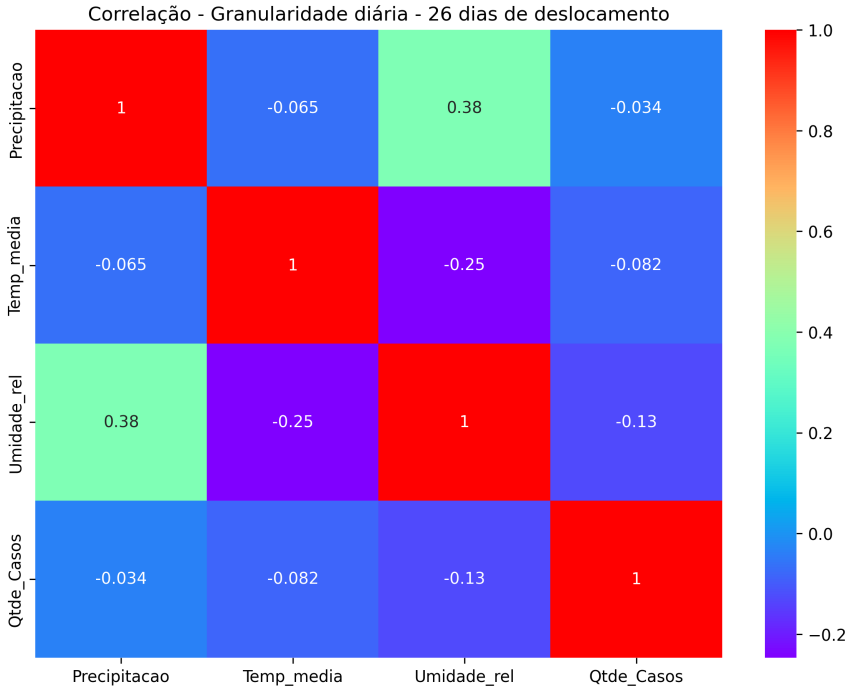
Fonte: o autor

Figura 53 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 25 dias



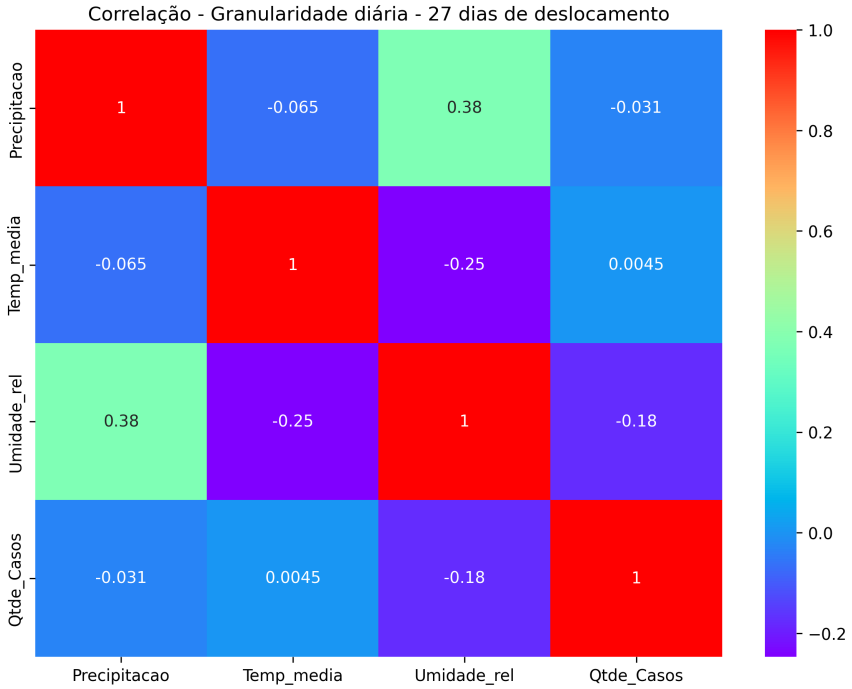
Fonte: o autor

Figura 54 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 26 dias



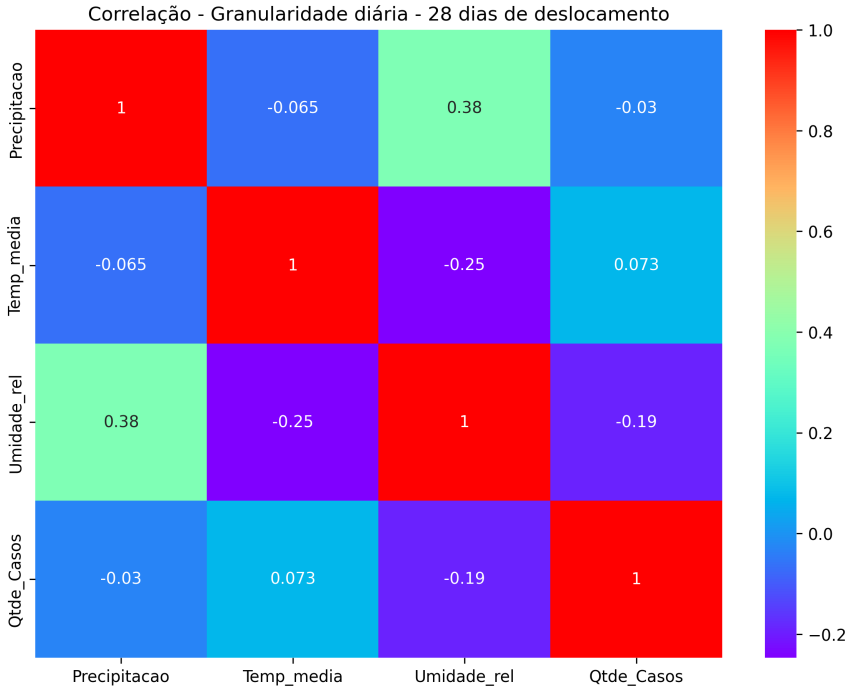
Fonte: o autor

Figura 55 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 27 dias



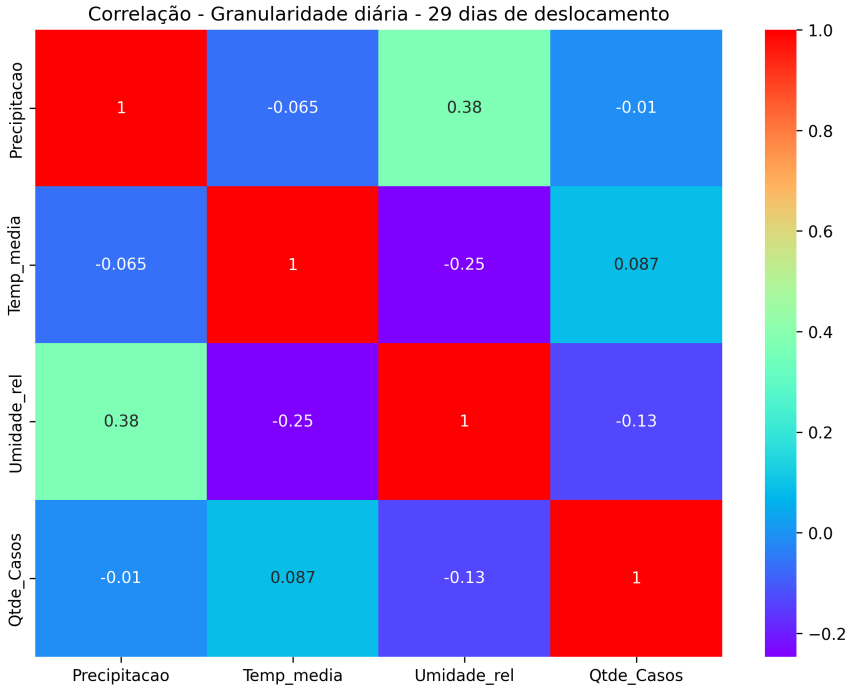
Fonte: o autor

Figura 56 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 28 dias



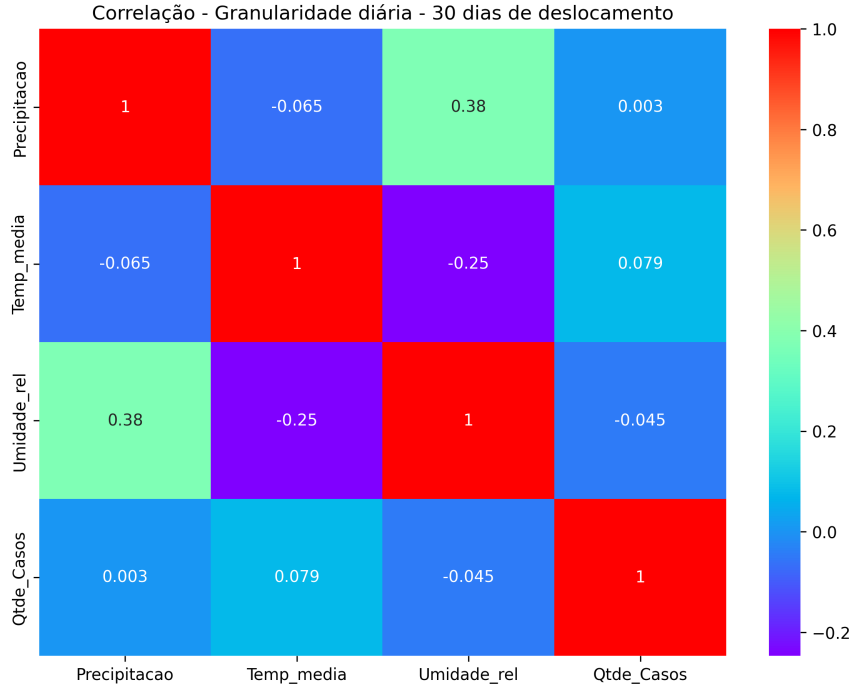
Fonte: o autor

Figura 57 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 29 dias



Fonte: o autor

Figura 58 – Matriz de correlação com granularidade diária com quantidade de casos deslocado em 30 dias



Fonte: o autor