



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
DEPARTAMENTO DE ENGENHARIAS E TECNOLOGIA
CURSO INTERDISCIPLINAR EM TECNOLOGIA DA INFORMAÇÃO

NATHAN CARDOSO LINHARES

**ANÁLISE COMPARATIVA ENTRE OS MODELOS ARIMA E LSTM APLICADOS À
MODELAGEM DE SÉRIES TEMPORAIS DE NOTIFICAÇÕES DE DENGUE NA
CIDADE DE MOSSORÓ**

PAU DOS FERROS

2025

NATHAN CARDOSO LINHARES

**ANÁLISE COMPARATIVA ENTRE OS MODELOS ARIMA E LSTM APLICADOS À
MODELAGEM DE SÉRIES TEMPORAIS DE NOTIFICAÇÕES DE DENGUE NA
CIDADE DE MOSSORÓ**

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Interdisciplinar em Tecnologia da Informação.

Orientador: George Felipe Fernandes Vieira, Prof.

PAU DOS FERROS

2025

© Todos os direitos estão reservados a Universidade Federal Rural do Semi-Árido. O conteúdo desta obra é de inteira responsabilidade do (a) autor (a), sendo o mesmo, passível de sanções administrativas ou penais, caso sejam infringidas as leis que regulamentam a Propriedade Intelectual, respectivamente, Patentes: Lei nº 9.279/1996 e Direitos Autorais: Lei nº 9.610/1998. O conteúdo desta obra tomar-se-á de domínio público após a data de defesa e homologação da sua respectiva ata. A mesma poderá servir de base literária para novas pesquisas, desde que a obra e seu (a) respectivo (a) autor (a) sejam devidamente citados e mencionados os seus créditos bibliográficos.

L735a Linhares, Nathan Cardoso.
Análise comparativa entre os modelos ARIMA e LSTM aplicados à modelagem de séries temporais de notificações de dengue na cidade de Mossoró / Nathan Cardoso Linhares. - 2025.
48 f. : il.

Orientador: George Felipe Fernandes Vieira.
Monografia (graduação) - Universidade Federal Rural do Semi-árido, Curso de Tecnologia da Informação, 2025.

1. LSTM. 2. ARIMA. 3. Dengue. 4. Série temporal. I. Vieira, George Felipe Fernandes, orient. II. Título.

Ficha catalográfica elaborada por sistema gerador automático em conformidade com AACR2 e os dados fornecidos pelo(a) autor(a).

Biblioteca Campus Pau dos Ferros
Bibliotecária: Erlanda Maria Lopes da Silva
CRB: 15/966

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

NATHAN CARDOSO LINHARES

**ANÁLISE COMPARATIVA ENTRE OS MODELOS ARIMA E LSTM APLICADOS À
MODELAGEM DE SÉRIES TEMPORAIS DE NOTIFICAÇÕES DE DENGUE NA
CIDADE DE MOSSORÓ**

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Interdisciplinar em Tecnologia da Informação.

Defendida em: 15/12/2025

BANCA EXAMINADORA

George Felipe Fernandes Vieira, Prof. (UFERSA)
Presidente

Amanda Gondim de Oliveira, Prof. dra. (UFERSA)
Membro Examinador

João Batista de Souza Neto, Prof. dr. (UFERSA)
Membro Examinador

Dedico este trabalho a Lauriana Cardoso e Sebastião Linhares, aos quais tenho a honra e o orgulho de chamar de pais.

AGRADECIMENTOS

Em meio a momentos de incerteza e profundo medo, nossa alma encontra-se em estado de vulnerabilidade. Felizmente, fui contemplado, ao longo da minha vida, com pessoas que me proporcionaram força e motivação para seguir em frente diante dos obstáculos.

Primeiramente, gostaria de agradecer aos meus pais e minha vó, que não mediram esforços e se dedicaram ao máximo para me proporcionar uma educação que eles não tiveram. Esse esforço é imensurável, e não consigo expressar em palavras o quanto sou grato por tê-los em minha vida.

Gostaria de agradecer a Kaylanne Alves, minha parceira de vida, cujas palavras serviram de conforto diante de momentos de caos. Sou extremamente grato por cada instante de tempo que compartilho ao seu lado.

Manifesto também minha gratidão ao professor e orientador George Felipe, por ter aceitado ser meu orientador e pela paciência e pelos valiosos ensinamentos transmitidos ao longo desta pesquisa. Sem dúvidas, sua contribuição foi fundamental para a realização deste trabalho.

Por último, mas não menos importante, sou imensamente grato aos meus colegas de curso, pois, sem o apoio de vocês ao longo da minha jornada, não teria sido possível alcançar este sucesso. Muito obrigado a todos.

“A existência é uma guerra e uma jornada em
terras desconhecidas.”

(Marco Aurélio)

RESUMO

As arboviroses constituem um relevante desafio para a saúde pública brasileira, com destaque para a dengue, doença transmitida pelo mosquito *Aedes aegypti* e amplamente disseminada em todo o território nacional. A compreensão da dinâmica temporal dessa enfermidade, bem como a capacidade de modelar e analisar seu comportamento histórico, são fundamentais para subsidiar ações de vigilância epidemiológica. Nesse contexto, este trabalho realiza uma análise comparativa entre dois modelos preditivos de naturezas distintas: um modelo estatístico clássico, o ARIMA, e uma rede neural recorrente baseada em Long Short-Term Memory (LSTM). Foram utilizados dados reais de notificações de dengue registradas na cidade de Mossoró, Rio Grande do Norte, Brasil, no período de 2010 a 2020. Após a realização do pré-processamento e da transformação dos dados em séries temporais univariadas, os modelos foram treinados e avaliados quanto à sua capacidade de aprender os padrões temporais históricos, utilizando as métricas RMSE, MAE e MAPE.

Palavras-chave: lstm; arima; dengue; série temporal;

ABSTRACT

Arboviral diseases represent a significant challenge to public health in Brazil, with dengue standing out as a widespread disease transmitted by the *Aedes aegypti* mosquito. Understanding the temporal dynamics of this disease, as well as the ability to model and analyze its historical behavior, is essential to support epidemiological surveillance actions. In this context, this study presents a comparative analysis between two predictive models of distinct natures: a classical statistical model, ARIMA, and a recurrent neural network based on Long Short-Term Memory (LSTM). Real dengue notification data recorded in the city of Mossoró, Rio Grande do Norte, Brazil, from 2010 to 2020 were used. After data preprocessing and transformation into univariate time series, the models were trained and evaluated regarding their ability to learn historical temporal patterns, using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) metrics.

Keywords: lstm; arima; dengue; time series;

LISTA DE FIGURAS

Figura 1 – Estrutura de um neurônio artificial	21
Figura 2 – Arquitetura de uma rede neural artificial	22
Figura 3 – Estrutura da camada LSTM	24
Figura 4 – Etapas da metodologia	32
Figura 5 – Total de notificações anuais de dengue entre os anos de 2010 e 2020	34
Figura 6 – Séries temporais das notificações de dengue no período de 2010 a 2020 no município de Mossoró	35
Figura 7 – Série temporal normalizada dos casos de dengue	37
Figura 8 – Comparativo dos resultados do treinamento do modelo LSTM, separados em treino, teste e predição	41
Figura 9 – Comparativo dos resultados do treinamento do modelo ARIMA, separados em treino, teste e predição	42
Figura 10 – Pseudo-código da lógica do grid search para o treinamento do modelo ARIMA	47

LISTA DE TABELAS

Tabela 1 – Dados referentes ao ano de 2010 da base utilizada neste estudo	33
Tabela 2 – Resultados das métricas RMSE, MAE e MAPE obtidos na fase de treinamento dos modelos LSTM e ARIMA.	40

LISTA DE QUADROS

Quadro 1 – Análise e comparação dos trabalhos relacionados	31
--	----

LISTA DE ABREVIATURAS E SIGLAS

R^2	<i>coefficient of determination</i>
AM	Aprendizado de Máquina
AR	Auto-regressivo
ARIMA	<i>Autoregressive Integrated Moving Average</i>
ES	<i>Exponential Smoothing</i>
FAC	Função de Autocorrelação
FACP	Função de Autocorrelação Parcial
GBR	<i>Gradient Boosting Regressor</i>
GRU	Gated Recurrent Unit
I	Integrado
LR	<i>Linear Regression</i>
LSTM	<i>Long Short Term Memory</i>
MA	Média Móvel
MAE	<i>Mean Absolute Error</i>
MAPE	<i>Mean Absolute Percentage Error</i>
MLP	<i>Multi layer Perceptron</i>
MSE	<i>Mean Squared Error</i>
NA	Neurônio Artificial
RFR	<i>Random Forest Regression</i>
RMSE	<i>Root Mean Squared Error</i>
RNA	Redes Neurais Artificiais
RNN	Rede Neural Recorrente
RW	<i>Random Walk</i>
SES	<i>Simple Exponential Smoothing</i>
SMOT	<i>Synthetic Minority Over-sampling TEchnique</i>
SVM	<i>Support vector machine</i>
SVR	<i>Support Vector Regression</i>
XGBoost	<i>Extreme Gradient Boosting</i>

SUMÁRIO

1	INTRODUÇÃO	15
2	FUNDAMENTAÇÃO TEÓRICA	18
2.1	Séries temporais	18
2.2	ARIMA	18
2.3	Aprendizado de Máquina	19
2.4	Neurônio Artificial e Redes Neurais Artificiais	20
2.5	Redes Neurais Recorrentes	22
2.6	LSTM	23
3	TRABALHOS RELACIONADOS	25
3.1	Uso de redes neurais recorrentes para predição de doenças isquêmicas do coração usando séries temporais	25
3.2	Comparação da performance de algoritmos de aprendizado de máquina para análise preditiva de febre amarela no estado de minas gerais	26
3.3	Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo aedes aegypti	28
3.4	Uma comparação entre métodos baseados em aprendizado de máquina para inferir número de casos semanais de dengue	29
3.5	Comparação entre os trabalhos relacionados	30
4	METODOLOGIA	32
4.1	Pré-processamento	32
4.1.1	Coleta de dados	32
4.1.2	Análise exploratória	33
4.2	Processamento	35
4.2.1	Treinamento	36
4.2.2	Treinamento do modelo LSTM	37

4.2.3	Treinamento do modelo ARIMA	39
5	RESULTADOS	40
5.1	Avaliação dos resultados	40
6	CONCLUSÕES E TRABALHOS FUTUROS	43
	REFERÊNCIAS	45
	APÊNDICE A EXEMPLO DE CÓDIGO	47

1 INTRODUÇÃO

Os arbovírus (*Arthropod-Borne Virus*), ou arboviroses, constituem um amplo grupo formado por centenas de vírus que compartilham a característica de serem transmitidos por artrópodes (Donalisio; Freitas; Zuben, 2017).

Segundo Donalisio, Freitas e Zuben (2017), as arboviroses vêm se consolidando como um problema crescente e persistente de saúde pública em escala mundial. Entre os principais fatores associados a esse cenário, destacam-se o elevado potencial de dispersão dos vetores, sua capacidade de adaptação a novos ambientes e hospedeiros, bem como o risco de ocorrência de epidemias extensas acompanhadas de casos graves.

As arboviroses transmitidas pelo mosquito *Aedes aegypti*, que atua como vetor ao disseminá-las por meio de sua picada, configuram-se como uma importante preocupação para a saúde pública (Silveira; Moreira, 2020).

Conforme destacado por Silveira e Moreira (2020), as arboviroses transmitidas pelo *Aedes aegypti* estão presentes em todas as unidades federativas do Brasil, sendo favorecidas por condições climáticas características dos períodos mais quentes e chuvosos, que propiciam a proliferação do mosquito, sobretudo em ambientes com acúmulo de água parada. No atual cenário epidemiológico brasileiro, um dos arbovírus de maior circulação é o DENV, agente etiológico da dengue (Donalisio; Freitas; Zuben, 2017).

Em 2019, o Brasil registrou 1.527.119 casos prováveis de dengue, correspondendo a uma taxa de incidência de 726,7 casos por 100 mil habitantes (BRASIL. Ministério da Saúde, 2019). A análise da distribuição regional desses casos evidencia expressivas diferenças epidemiológicas entre as regiões do país. Nesse contexto, a Região Nordeste apresentou uma taxa de incidência de 372,8 casos por 100 mil habitantes, valor inferior à média nacional, porém indicativo de significativa circulação viral (BRASIL. Ministério da Saúde, 2019). Esses dados demonstram que, embora o Nordeste não tenha figurado entre as regiões com maior incidência no período analisado, a dengue manteve-se como um relevante problema de saúde pública regional.

No contexto local, essa problemática também se manifesta de forma expressiva. No município de Mossoró, localizado no estado do Rio Grande do Norte, região Nordeste do Brasil, a Vigilância Epidemiológica Municipal notificou, em 2020, a confirmação de 1.202 casos de dengue, 196 casos de zika e 2.262 casos de chikungunya (Vieira; Castro; Oliveira, 2023). Esses dados evidenciam a relevância das arboviroses para a saúde pública municipal.

De acordo com MELO *et al.* (2025), apesar da redução no número de casos observada

durante os períodos mais críticos da pandemia de COVID-19, possivelmente associada às mudanças nos padrões de mobilidade populacional e às medidas de isolamento social, a retomada das atividades regulares a partir de 2022 sinaliza um provável aumento na incidência das arboviroses.

Diante da magnitude desse cenário epidemiológico, torna-se evidente a necessidade de instrumentos capazes de auxiliar na compreensão, monitoramento e antecipação do comportamento das arboviroses.

Sob a perspectiva das inovações e dos avanços tecnológicos, a área da saúde está entre as mais beneficiadas pela tecnologia (Nilsson, 2014). Desde os anos 2000, diversos estudos vêm sendo desenvolvidos com o uso de Redes Neurais, algoritmos de Aprendizado de Máquina e modelos estatísticos, com o intuito de contribuir para o monitoramento e a previsão da ocorrência de casos de dengue (Andrade; Brilhante, 2018).

Em razão das crescentes preocupações relacionadas à dengue e às demais arboviroses, Silveira e Moreira (2020) destacam a urgência na adoção de medidas eficazes não apenas no combate, mas, principalmente, na prevenção dessas doenças.

Nesse contexto, modelos preditivos configuram-se como ferramentas essenciais, uma vez que possibilitam a antecipação de potenciais surtos e subsidiam decisões estratégicas relacionadas à alocação de recursos humanos, materiais e financeiros (Silveira; Moreira, 2020). Ao analisar padrões históricos e comportamentos epidemiológicos, métodos baseados em inteligência computacional podem auxiliar gestores e profissionais de saúde na detecção precoce de situações críticas e no planejamento de ações preventivas mais eficientes (Silveira; Moreira, 2020).

Todavia, conforme apontado por Correia *et al.* (2021), o desenvolvimento de modelos preditivos capazes de produzir previsões confiáveis em séries temporais constitui um dos principais desafios atuais. Isso se deve, sobretudo, à alta dimensionalidade dos dados, à dependência entre as observações e à complexidade dos comportamentos epidemiológicos (Correia *et al.*, 2021).

Portanto, o presente trabalho tem como objetivo avaliar e comparar o desempenho e a eficiência de modelos preditivos de natureza estatística e baseada em aprendizado profundo na previsão de casos de dengue. Para tanto, são aplicados os modelos *Autoregressive Integrated Moving Average* (ARIMA) e *Long Short Term Memory* (LSTM) a uma série temporal composta por notificações reais de dengue referentes ao município de Mossoró. A análise comparativa

é conduzida a partir de métricas quantitativas de avaliação, onde as métricas são: *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE) e *Mean Absolute Percentage Error* (MAPE). Logo, permitem identificar a abordagem que apresenta maior capacidade preditiva no cenário estudado.

O presente trabalho está estruturado da seguinte forma: o capítulo 2 apresenta a fundamentação teórica, contemplando os principais conceitos necessários para a compreensão do estudo. No capítulo 3, são discutidos os trabalhos relacionados que compactuam com a pesquisa. O capítulo 4 descreve a metodologia adotada, detalhando os procedimentos utilizados para a execução e organização da pesquisa. Em seguida, o capítulo 5 expõe os resultados obtidos, bem como a análise e discussão desses achados. Por fim, o Capítulo 6 reúne as conclusões do estudo e sugere possíveis direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é abordado um conjunto de conceitos na área de estudo do projeto, onde, diante esses conceitos, pretende-se ter a finalidade de uma melhor compreensão do presente trabalho.

2.1 Séries temporais

O presente estudo utilizou as séries temporais para realizar a modelagem dos casos de dengue. Esta abordagem, é amplamente utilizada para observar a evolução de uma variável ao longo do tempo, possibilitando captar possíveis padrões. Assim, essa modelagem pode agregar no treinamento dos modelos ARIMA e LSTM, pois, com as séries temporais, ambos os modelos podem identificar eventuais padrões e considera-los em suas predições.

Uma série temporal pode ser caracterizada como um conjunto de observações de uma variável aleatória, disposta e organizada em sequências de tempos. Geralmente, tais observações são equiespaçadas, ou seja, estão dispostas em uma mesma frequência de tempo como ano, mês, semana, dias, horas e etc. (Brockwell; Davis, 2002).

Uma forma de representar, de modo teórico, uma série temporal é considerando uma variável aleatória Y no instante de tempo t , na qual a série temporal é escrita da seguinte maneira: $Y_1, Y_2, Y_3, \dots, Y_n$, n é o número de observações da variável aleatória no intervalo de t . (SOUZA; CAMARGO, 1996).

A decomposição clássica das séries temporais, onde estão presentes quatro principais componentes que, através destes, são realizadas as previsões. Tais componentes são: tendência, refere-se a direção geral que é apresentada na série temporal conforme é desenvolvida no intervalo de tempo; sazonalidade, diz respeito a padrões previsíveis que aparecem com regularidade e em períodos de tempo curto e fixos; ciclos, são variações de longo prazo que são representados em períodos mais extensos do que a sazonalidade, não necessariamente são regulares ou previsíveis; por fim, o termo aleatório, que são flutuações de curto prazo com deslocamento inexplicável (Brockwell; Davis, 2002).

2.2 ARIMA

O modelo ARIMA foi proposto por Box e Jenkins (1976), com o objetivo de tentar representar matematicamente as variações presentes em séries temporais. A construção dos

modelos ARIMA ocorre por meio de um processo iterativo, no qual a definição da estrutura do modelo é baseada pelos próprios dados (Box; Jenkins, 1976).

O modelo de Box e Jenkins é composto por três partes principais que explicam sua estrutura. A porção Auto-regressivo (AR) indica que a variável de interesse é regressada em seus próprios valores anteriores. A componente Média Móvel (MA) mostra que o erro do modelo depende de combinações lineares de erros passados e atuais, refletindo como imprecisões anteriores influenciam previsões futuras. Por fim, a parte Integrado (I) que diz respeito a transformação dos dados por meio da diferenciação dos valores atuais com os valores passados (Box; Jenkins, 1976).

No estudo desenvolvido por (Souza *et al.*, 2024), é abordado os três passos para a realização das previsões, utilizando o modelo ARIMA, assim como o funcionamento dos seus parâmetros (p, d, q) , sendo os seguintes passos:

1. **Identificação de estacionariedade:** consiste em verificar se a série temporal é estacionária, caso contrário, realiza-se a diferenciação dos dados até que a série atinja a estacionariedade, sendo o número de diferenciações necessário representado pelo parâmetro d ;
2. **Estimação e teste:** são determinados os valores apropriados para os parâmetros p e q do modelo ARIMA, onde são definidos a partir da análise dos gráficos das Função de Autocorrelação (FAC) e de Função de Autocorrelação Parcial (FACP), considerando os atrasos da série, essa estimativa pode ser realizada por meio de mínimos quadrados ou outros métodos;
3. **Previsão:** passo final que resulta, diante os passos anteriores, em um modelo que é utilizado para realizar previsões.

Para este estudo, a escolha do modelo ARIMA se justifica pelo motivo de ser uma abordagem clássica em relação a previsões de séries temporais univariadas, ou seja, uma série temporal que consiste em uma única variável. Logo, nesta pesquisa, as previsões utilizam apenas as observações de casos de dengue ao longo dos anos, sendo assim, o modelo ARIMA, uma abordagem coerente para a previsão dos casos.

2.3 Aprendizado de Máquina

Catarino (2025) afirma que Aprendizado de Máquina (AM) é um subconjunto da área de *Inteligência Artificial* que visa técnicas computacionais com o propósito de máquinas aprenderem e evoluírem sem a necessidade de serem programadas. Também é afirmado por (Paixão *et al.*,

2022) que AM tenta buscar uma interseção de técnicas da área da matemática e estatística com algoritmos computacionais.

Os modelos de AM tem como objetivo criar um sistema computacional que aprenda com base em um banco de dados pré-definido e, ao final, o modelo possa realizar previsões, detecções ou classificações (Paixão *et al.*, 2022).

No trabalho realizado por (Paixão *et al.*, 2022), é apresentado o processo de desenvolvimento do algoritmo de aprendizado de máquina, que é composto por três fases:

- **Pré-processamento:** consiste na preparação do algoritmo por meio da organização do banco de dados, definição da pergunta de pesquisa e divisão dos dados entre treinamento e teste;
- **Treinamento:** é o processo em que o algoritmo aprende com base nos métodos de aprendizado supervisionado, onde o modelo é treinado a partir de um conjunto de dados rotulado, ou não supervisionado, os dados não possuem rótulos, e o algoritmo busca identificar padrões e relações ocultas entre as informações, como agrupamentos ou associações;
- **Avaliação do modelo:** etapa final em que o modelo é comparado com os dados de teste, e os resultados são analisados para verificar sua eficiência.

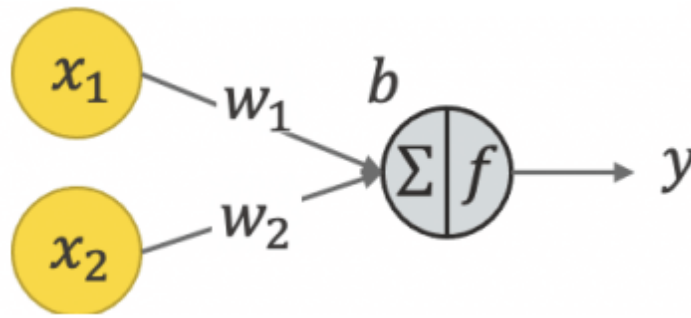
A utilização de técnicas de Aprendizado de Máquina possibilita que os modelos aprendam a partir dos dados e identifiquem padrões neles presentes. Nesse sentido, este trabalho emprega técnicas de aprendizado supervisionado, nas quais os dados referentes ao número de notificações de dengue são utilizados no processo de treinamento dos modelos e posteriormente comparados com os dados de teste para fins de avaliação.

2.4 Neurônio Artificial e Redes Neurais Artificiais

As redes neurais artificiais foram utilizadas neste trabalho por sua capacidade de aprender diretamente com os dados e se adaptar aos padrões observados. Os modelos que utilizam Redes Neurais Artificiais (RNA) têm a capacidade de ajustar automaticamente seus parâmetros durante o processo de treinamento. Assim, com a utilização de RNA, possibilitou-se uma abordagem de treinamento e predição diferente dos métodos estatísticos clássicos.

Na década de 1940, foi proposto por McCulloch e Pitts, o modelo matemático para o neurônio artificial, tal modelo que foi fundamental para o desenvolvimento das RNAs (Souza *et al.*, 2025). O modelo consiste em sinais de entrada, pesos sinápticos, um combinador linear e uma função de ativação (Souza *et al.*, 2025).

Figura 1 – Estrutura de um neurônio artificial



Fonte: Adaptado de Nei Grando (2022).

É apresentado por Souza *et al.* (2025) que o neurônio artificial recebe um conjunto de entradas $x_1, x_2, x_3, \dots, x_n$, as quais são multiplicadas pelos respectivos pesos sinápticos $w_1, w_2, w_3, \dots, w_n$ e somadas ao viés b . O resultado do somatório Σ passa, então, por uma função de ativação f , que transforma a soma obtida em uma saída y . Essa saída determina se o neurônio será ativado ou não. Diante da Figura 1, pode-se observar a representação visual de um Neurônio Artificial (NA).

A partir da combinação e interconexão de múltiplos neurônios artificiais organizados em camadas, forma-se uma estrutura capaz de modelar relações complexas entre os dados de entrada e saída, dando origem às Redes Neurais Artificiais (Silva; Spatti; Flauzino, 2010).

As RNA são modelos computacionais inspirados no sistema nervoso dos seres vivos (Silva; Spatti; Flauzino, 2010). Possuem a capacidade de aquisição e manutenção do conhecimento, baseando-se em informações, e podem ser definidas como um conjunto de unidades de processamento, caracterizadas por neurônios artificiais, que são interligados por um grande número de interconexões (*sinapses artificiais*), sendo representadas aqui por vetores/matrizes de pesos sinápticos (Silva; Spatti; Flauzino, 2010).

As RNA possuem arquiteturas que são a estrutura que define a forma como os seus diversos NA estão arranjados, ou dispostos, uns em relação aos outros, na qual esses arranjos são essencialmente estruturados através do direcionamento das conexões sinápticas dos neurônios (Silva; Spatti; Flauzino, 2010).

Segundo Silva, Spatti e Flauzino (2010), uma RNA pode ser dividida em três partes, normalmente conhecidas de camadas, as quais são denominadas da seguinte forma:

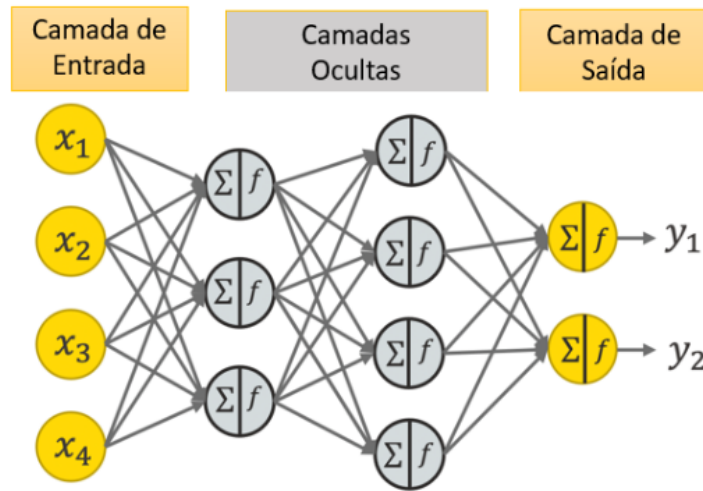
- Camada de Entrada (*Input layer*): é a camada responsável pelo recebimento de informações, ou seja, os dados, sinais, características ou medições advindas do meio externo;
- Camada Oculta (*Hidden layer*): quase todo o processamento interno da rede é realizado nesta camada, pois, é composta por neurônios que possuem a responsabilidade de extrair

as características associadas ao processo ou sistema que é inserido;

- Camada de Saída (*Output layer*): esta camada é constituída por neurônios que se responsabilizam pela produção e apresentação dos resultados finais da rede.

É apresentado pela Figura 2, uma ilustração da Camada de Entrada, Camada Oculta e Camada de Saída de uma RNA.

Figura 2 – Arquitetura de uma rede neural artificial



Fonte: Adaptado de Nei Grando (2022).

2.5 Redes Neurais Recorrentes

As *Redes Neurais Recorrentes* (RNN) foram criadas para processar sequências de dados em que a ordem dos elementos tem grande importância, por exemplo, séries temporais, textos, áudios e etc. Sendo sua principal característica possuir conexões recorrentes, tais conexões, permitem com que a informação seja mantida e reutilizada ao longo do tempo (Carvalho; Cortes, 2025).

Segundo (Catarino, 2025), as RNNs conseguem armazenar as entradas de dados de forma sequencial em um histórico, criando conexões entre os neurônios artificiais que geram ciclos, esses ciclos correspondem às conexões recorrentes da rede.

De acordo com (Carvalho; Cortes, 2025), a formulação fundamental de uma Rede Neural Recorrente (RNN) pode ser expressa pela seguinte equação: $h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b)$, na qual o σ representa a função de ativação responsável por introduzir não linearidade no modelo; h_t é o estado oculto no instante de tempo t ; W_h corresponde aos pesos associados ao estado oculto anterior (h_{t-1}); W_x representa os pesos aplicados às entradas atuais (x_t); e b é o termo de *bias*, ou

seja, o viés que ajusta a saída do neurônio, permitindo maior flexibilidade no aprendizado.

2.6 LSTM

O modelo LSTM difere das RNN convencionais por conter blocos de memórias que se encontram conectados de forma recorrente, onde esses blocos são compostos por células que conseguem armazenar o estado temporal da rede neural (Catarino, 2025).

A cada passo ao longo do tempo, a RNN produz uma saída com base no estado oculto, de acordo com a expressão: $y_t = \sigma(W_y \cdot h_t + b_y)$. A partir disso, observa-se que as RNNs convencionais possuem “memória” de curto prazo, o que as torna adequadas para certas tomadas de decisão. No entanto, elas não conseguem capturar tendências de longo prazo devido ao problema de explosão ou desaparecimento do gradiente. Diante esse problema do gradiente, foi criado o LSTM, um modelo com a proposta de solucionar o problema de memória de curto prazo (Carvalho; Cortes, 2025).

De modo a compreender o funcionamento do LSTM, (Carvalho; Cortes, 2025) explica o sistema de *gates* do modelo, que diz respeito à forma como ele resolve o problema da memória de longo prazo. Esse mecanismo utiliza três tipos de *gates* (portões, em tradução literal) responsáveis por controlar o fluxo de informações: o *forget gate* (portão do esquecimento), que decide quais informações devem ser descartadas; o *input gate* (portão de entrada), que determina quais novas informações serão armazenadas; e o *output gate* (portão de saída), que define quais informações serão expostas como saída. Além disso, as LSTM possuem uma célula de memória capaz de armazenar informações a longo prazo. A Gated Recurrent Unit (GRU), por sua vez, é uma variação simplificada das LSTM, composta por dois *gates*: o *reset gate*, que controla quais informações passadas devem ser esquecidas, e o *update gate*, que decide quanto do estado anterior deve ser mantido.

Na Figura 3, é ilustrado o funcionamento da camada do LSTM e seus *gates*. (Ribeiro, 2024) sintetiza a imagem descrevendo que C_{t-1} corresponde ao estado da unidade de memória no instante anterior. A unidade de memória é o componente principal da LSTM e é usada para controlar a transmissão da informação. O estado da camada no instante anterior é representado por h_{t-1} e a sequência de entrada no instante atual representado por x_t . O portão de entrada, o portão de esquecimento e portão de saída são representados respectivamente por i , f e o . Logo, i_t , f_t e o_t representam, respectivamente as equações: $\sigma(W_i[h_{t-1}, x_t] + b_i)$, $\sigma(W_f[h_{t-1}, x_t] + b_f)$ e $\sigma(W_o[h_{t-1}, x_t] + b_o)$. As matrizes W_i , W_f e W_o são as matrizes de pesos dos portões de entrada,

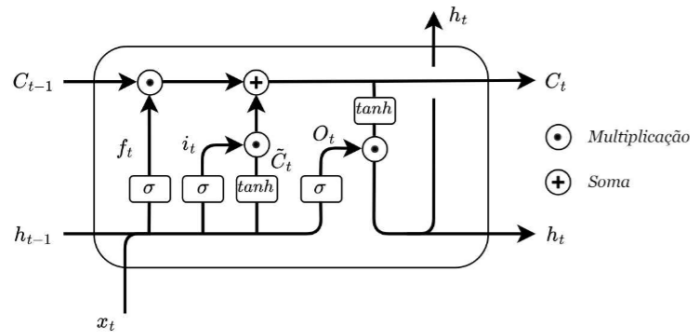
esquecimento e saída, de forma respectiva. Os vetores de vieses b_i , b_f e b_o correspondem os vieses de seus respectivos portões, onde as matrizes e vieses equivalentes são aplicados na função de ativação sigmoid, representada pelo símbolo σ (Ribeiro, 2024).

Por último, o novo estado da camada h_t é obtido diante as equações:

- $\tilde{C}_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$
- $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$
- $h_t = o_t \odot \tanh(C_t)$

Na qual o símbolo $\odot$ representa a multiplicação termo a termo dos vetores e $\tanh$ é a função de ativação tangente (Ribeiro, 2024).

Figura 3 – Estrutura da camada LSTM



Fonte: Adaptado de Ribeiro (2024).

Diante dos estudos que abordam o modelo LSTM, observa-se sua relevância no uso de RNN para problemas que envolvem a previsão de dados futuros. Além de ser amplamente empregado em tarefas de predição, o LSTM mostra-se um modelo coerente para a análise dos casos de dengue, permitindo a realização de um comparativo adequado com o modelo ARIMA, pois ambos os modelos tratam de abordagens amplamente utilizadas.

3 TRABALHOS RELACIONADOS

Neste capítulo, serão discutidos e apresentados os trabalhos que compactuam com o presente estudo. Na seção 3.1, será abordado o uso de Redes Neurais Recorrentes na predição de doenças. Na seção 3.2, será apresentada a comparação de performance entre modelos de Aprendizado de Máquina com objetivos preditivos. A seção 3.3 mostrará a utilização de algoritmos de Aprendizado de Máquina para a predição de arboviroses, especificamente, doenças transmitidas pelo *Aedes aegypti*. Por fim, na seção 3.4, será apresentada a comparação entre modelos estatísticos e algoritmos de Aprendizado de Máquina aplicados para predição de casos de dengue.

3.1 Uso de redes neurais recorrentes para predição de doenças isquêmicas do coração usando séries temporais

No que diz respeito à utilização de Redes Neurais Recorrentes aplicadas a séries temporais para a predição de doenças, destaca-se a sua versatilidade em explorar séries temporais com diferentes variáveis no âmbito da saúde. O trabalho de (Carvalho; Cortes, 2025) explorou o uso das RNNs para a predição de doenças isquêmicas na cidade de São Luís do Maranhão.

A pesquisa de (Carvalho; Cortes, 2025) buscou avaliar a eficiência dos modelos de RNNs aplicados a uma série temporal dos números de óbitos causados por doenças isquêmicas do coração, como infarto e angina, na cidade de São Luís do Maranhão. Os modelos utilizados nessa pesquisa foram o LSTM, BiLSTM e GRU, com o objetivo de identificar quais parâmetros resultariam em melhores desempenhos. Para realizar a busca dos melhores parâmetros, foi utilizado uma separação nos dados de treinamento e teste, sendo 80% destinado ao treinamento e 20% aos teste. Logo após a separação, foi realizado o treinamento dos modelos e aplicado a técnica de validação cruzada de k-fold, possibilitando testar diferentes subconjuntos de treinamento e teste. Por fim, os modelos em questão foram comparados com o modelo estatístico ARIMA, por meio das métricas RMSE e MAPE, permitindo avaliar o desempenho da predição dos modelos.

De acordo com (Carvalho; Cortes, 2025), ao analisar as métricas de desempenho, observou-se que os modelos baseados em RNN apresentaram resultados significativamente superiores. O modelo que obteve o melhor desempenho foi o BiLSTM, alcançando um RMSE de 0,45 e um MAPE de 0,010. Os demais modelos apresentaram os seguintes resultados, por ordem de melhor desempenho: GRU com RMSE igual a 0,46 e MAPE de 0,010; LSTM com

RMSE de 0,70 e MAPE de 0,16; enquanto o modelo ARIMA apresentou o pior desempenho, com RMSE de 7,5 e MAPE de 0,11.

Assim, (Carvalho; Cortes, 2025) concluiu em seu trabalho que o modelo BiLSTM apresentou o melhor resultado, ainda que de forma sutil em comparação com os demais modelos de RNN. Entretanto, foi observado que a proximidade entre os resultados pode indicar que a bidirecionalidade não é um fator determinante. Ademais, (Carvalho; Cortes, 2025) também ressalta que, devido à baixa complexidade do problema, o uso do modelo BiLSTM pode ser desnecessário, uma vez que ele requer maior poder computacional em relação aos demais.

Em comparativo com o trabalho realizado por (Carvalho; Cortes, 2025), no qual foram utilizados os mesmos modelos e métricas semelhantes a presente pesquisa, porém, para a variável de óbitos por doenças isquêmicas do coração, o estudo contribuiu ao demonstrar a relevância desses modelos para a predição de séries temporais com diferentes variáveis. Além de apresentar uma comparação direta entre eles. Entretanto, o presente trabalho distingue-se ao aplicar a modelagem de séries temporais para a previsão de casos de dengue, explorando um contexto epidemiológico e uma variável distinta a (Carvalho; Cortes, 2025). Ademais, não foram empregadas as arquiteturas BiLSTM, GRU e, embora a separação dos dados entre treino e teste siga a mesma abordagem adotada por (Carvalho; Cortes, 2025), este estudo não utilizou a técnica de validação cruzada do tipo K-Fold na etapa posterior ao treinamento dos modelos.

3.2 Comparação da performance de algoritmos de aprendizado de máquina para análise preditiva de febre amarela no estado de minas gerais

Ao passo em que modelos de aprendizado de máquina surgem com o propósito preditivo, tem-se a possibilidade de compará-los e medir tamanha eficiência em suas predições. Tomando como exemplo, o trabalho de (Araújo *et al.*, 2023) abordou a comparação de performance de algoritmos de aprendizado de máquina para análise preditiva dos casos de febre amarela no estado de Minas Gerais, Brasil.

A tese de (Araújo *et al.*, 2023) tem como objetivo utilizar técnicas supervisionadas de AM e comparar a performance em problemas de regressão para prever a ocorrência de febre amarela humana em Minas Gerais. (Araújo *et al.*, 2023) condensou sua pesquisa de forma a estrutura-la em três artigos complementares: o primeiro consistiu em uma revisão integrativa da literatura sobre a ecoepidemiologia da febre amarela na região Sudeste do Brasil, que subsidiou a seleção inicial das variáveis analisadas; o segundo apresentou um estudo epidemiológico retrospectivo e

descritivo dos dados do SINAN referentes aos casos humanos de febre amarela e às epizootias em primatas não humanos registrados em Minas Gerais entre 2007 e 2020; e o terceiro desenvolveu e comparou modelos preditivos para estimar a ocorrência de casos humanos da doença no estado, avaliando o desempenho das diferentes abordagens de modelagem propostas.

No âmbito da presente pesquisa, o terceiro artigo desenvolvido por (Araújo *et al.*, 2023) apresenta maior proximidade com este estudo, uma vez que explora a aplicação de modelos de AM para a predição de casos de febre amarela observados em séries temporais e, posteriormente, avalia o desempenho dessas abordagens, alinhando-se a proposta metodológica.

No terceiro artigo, (Araújo *et al.*, 2023) avaliou um conjunto de algoritmos supervisionados de regressão com o propósito de estimar a ocorrência de febre amarela humana a partir de dados epidemiológicos e ambientais. Foi utilizado quatro modelos de AM que são: *Random Forest Regression* (RFR), *Extreme Gradient Boosting* (XGBoost), *Multi layer Perceptron* (MLP) e *Support vector machine* (SVM), todos submetidos a processos de ajuste e otimização de hiperparâmetros, o desempenho de cada modelo foi mensurado utilizando métricas RMSE, MAE, *Mean Squared Error* (MSE) e *coefficient of determination* (R^2).

(Araújo *et al.*, 2023) observou em seus resultados que random forest obteve o melhor desempenho, diante sua abordagem comparativa, apresentando MSE igual 0,22; MAE igual 0,07; RMSE igual 0,47; e R^2 igual 77,68% com pouca diferença de performance para a rede MLP e XGBoost. Por fim, (Araújo *et al.*, 2023) concluiu que os modelos de previsão da febre amarela utilizando algoritmos de aprendizado de máquina com aprendizado supervisionado apresentaram ser boas ferramentas para previsão de série temporal multivariada.

O trabalho de (Araújo *et al.*, 2023) apresenta uma contribuição relevante ao propor uma abordagem robusta para a previsão de séries temporais, utilizando modelos supervisionados de aprendizado de máquina aplicados aos casos de febre amarela em Minas Gerais. Contudo, o estudo explorou uma ampla gama de algoritmos de regressão multivariada, direcionando seu foco principal para a comparação de desempenho entre diferentes modelos de AM. Em contraste, o presente trabalho concentra-se de forma mais específica em apenas um modelo de aprendizado de máquina, o LSTM, além de um modelo estatístico tradicional, o ARIMA, aprofundando-se na análise comparativa entre essas duas abordagens.

3.3 Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo *aedes aegypti*

No contexto da aplicação de algoritmos de AM na predição de arboviroses, observa-se a adoção de diferentes estratégias metodológicas para estimar a ocorrência de arboviroses. O trabalho de (Silveira; Moreira, 2020) empregou técnicas de inteligência computacional com o objetivo de prever a existência de dengue e chikungunya a partir de características clínicas e demográficas dos pacientes.

Segundo (Silveira; Moreira, 2020), foram utilizados algoritmos de aprendizado de máquina para prever casos das duas arboviroses com base em atributos como sintomas, idade, sexo, período de manifestação dos sintomas e demais informações contidas nas fichas de notificação. O estudo, realizado na cidade de Recife, Pernambuco, Brasil, adotou algoritmos de classificação, especificamente o J48, o *Random Forest* e o *Multilayer Perceptron*. Para avaliar o desempenho dos modelos, foram utilizadas métricas como acurácia, *precision*, *recall* e *f-measure*.

O estudo de (Silveira; Moreira, 2020), adotou uma estrutura de três etapas na sua metodologia, são essas etapas, respectivamente: pré-processamento, processamento e análise. Na etapa de pré-processamento, (Silveira; Moreira, 2020) realizou o processo de coletar dados de dengue, chikungunya e outras doenças, realizando a limpeza e padronização dos atributos a fim de eliminar ruídos e garantir a consistência do conjunto de dados. Após ajustes e integração das bases, foram mantidos 42 atributos considerados relevantes para a tarefa preditiva.

Durante a etapa de processamento, houve a necessidade de balancear as classes, uma vez que havia forte predominância de casos de dengue em relação às demais categorias. Para resolver esse desbalanceamento, utilizou-se a técnica *Synthetic Minority Over-sampling TEchnique* (SMOT), que gera exemplos sintéticos da classe minoritária. No estudo, aplicou-se o SMOT progressivo, iniciando em 100% e variando até 700%, além de combinações de duas aplicações sucessivas da técnica, com o objetivo de analisar o impacto do balanceamento nas métricas de avaliação.

Na etapa de análise, após a execução dos modelos J48, *Random Forest* e *Multilayer Perceptron*, (Silveira; Moreira, 2020) compararam os resultados obtidos por meio das métricas de *precision*, *recall*, acurácia, *f-measure*, matriz de confusão e tempo de construção dos modelos.

(Silveira; Moreira, 2020) concluiu, diante os resultados apresentados, que o algoritmo *Random Forest* apresentou o melhor desempenho geral, alcançando 90,64% de acurácia e valor de *f-measure* de 0,907, especialmente em cenários onde o conjunto de dados foi totalmente

balanceado pelo SMOTE.

O trabalho de (Silveira; Moreira, 2020) apresenta uma contribuição relevante ao tratar diretamente da predição de arboviroses, especificamente, dengue e chikungunya. Contudo, sua abordagem difere do presente estudo na utilização de técnicas de modelagem diferente das séries temporais. (Silveira; Moreira, 2020), baseia-se em dados clínicos e demográficos dos pacientes, caracterizando-se, a predição mediante um problema de classificação. Em razão dessa distinção do conjunto de dados, foram empregados modelos diferentes do LSTM e ARIMA. Outro ponto diferente deste trabalho, é a utilização da técnica específica de balanceamento de classes, o SMOT.

3.4 Uma comparação entre métodos baseados em aprendizado de máquina para inferir número de casos semanais de dengue

Mediante a aplicação de modelos preditivos voltados à análise de casos de dengue, pode-se explorar uma ampla gama de abordagens, sejam elas estatísticas ou baseadas em Redes Neurais. No entanto, o trabalho de (Zanardo *et al.*, 2024) contemplou um número considerável de modelos, trazendo abordagens de modelos estatísticos clássicos, redes rasas e aprendizado profundo, utilizando dados referentes ao número de casos de dengue na cidade de San Juan, Porto Rico.

De acordo com o estudo de (Zanardo *et al.*, 2024), os modelos foram divididos em três categorias: modelos estatísticos clássicos, como o *Random Walk* (RW), o modelo linear de Holt, *Simple Exponential Smoothing* (SES), o modelo de Holt-Winters ou *Exponential Smoothing* (ES) e o ARIMA; modelos rasos regressivos, incluindo *Linear Regression* (LR), *Support Vector Regression* (SVR), RFR, *Gradient Boosting Regressor* (GBR) e XGBoost; e, por fim, arquiteturas de aprendizado profundo, como CNN-LSTM e LSTNet. Além disso, os dados não foram tratados diretamente como séries temporais, mas transformados em uma matriz de correlação cruzada, sobre a qual foram aplicados filtros de transformação *wavelet*.

O experimento descrito por (Zanardo *et al.*, 2024) aplicou todos os modelos a matriz de correlação cruzada contendo os casos semanais de dengue da cidade de San Juan e realizou uma comparação com base nas métricas RMSE, MAE, MAPE, R^2 e Viés Médio (*BIAS*).

Os resultados obtidos por (Zanardo *et al.*, 2024) evidenciam que o modelo de Regressão Linear, combinado com a filtragem pela *wavelet sym20*, apresentou o melhor desempenho geral entre todos os métodos avaliados, alcançando valores reduzidos de RMSE e MAPE quando

comparado aos demais. Apesar de modelos mais complexos, como o CNN-LSTM-LR, também demonstrarem bons resultados, o estudo conclui que o aumento da complexidade nem sempre se traduz em melhor capacidade preditiva, especialmente em conjuntos de dados com menor volume e variabilidade. Ademais, o trabalho destaca que a Regressão Linear se mostrou mais eficiente por sua simplicidade e menor custo computacional, sendo suficiente para capturar as relações lineares presentes nos dados de San Juan.

O estudo de (Zanardo *et al.*, 2024) apresenta contribuições significativas ao comparar um amplo conjunto de modelos destinados a inferência de casos de dengue. Todavia, o presente trabalho, distingue-se no aspecto em que não foram explorados modelos de aprendizado raso, limitando-se a apenas dois métodos preditivos. Além disso, (Zanardo *et al.*, 2024) empregou uma modelagem baseada na transformação dos dados em uma matriz de correlação cruzada ao invés da modelagem de séries temporais. Outra distinção relevante diz respeito ao conjunto de métricas de avaliações adicionais, como R^2 e *Bias*. (Zanardo *et al.*, 2024), por fim, trouxe maior robustez em investigar arquiteturas híbridas mais complexas, como o modelo CNN-LSTM-LR, as quais não foram adotadas nesta pesquisa.

3.5 Comparação entre os trabalhos relacionados

Com base na análise dos trabalhos relacionados apresentados e sintetizados no Quadro 1, observa-se que a literatura recente tem explorado de forma recorrente o uso de técnicas computacionais aplicadas à área da saúde, especialmente no contexto da predição de doenças a partir de dados históricos. Entretanto, tais estudos diferem significativamente quanto às abordagens metodológicas adotadas, aos tipos de dados analisados e aos objetivos preditivos.

Os trabalhos de Carvalho e Cortes (2025) e Araújo *et al.* (2023) destacam-se pela utilização de séries temporais no contexto epidemiológico, porém direcionam suas análises para doenças distintas da dengue. Enquanto Carvalho e Cortes (2025) realiza uma comparação entre modelos baseados em redes neurais recorrentes e o modelo estatístico ARIMA, seu foco recai sobre óbitos por doenças isquêmicas do coração. Por outro lado, Araújo *et al.* (2023) concentra-se na comparação de algoritmos supervisionados de aprendizado de máquina para a predição de febre amarela, sem empregar modelos recorrentes ou estatísticos clássicos como o ARIMA.

O estudo de Silveira e Moreira (2020), embora trate diretamente da predição de arboviroses, adota uma abordagem distinta ao formular o problema como uma tarefa de classificação

baseada em dados clínicos e demográficos, não explorando a modelagem de séries temporais nem o uso dos modelos LSTM ou ARIMA. Já o trabalho de Zanardo *et al.* (2024) apresenta uma análise abrangente ao comparar múltiplos modelos estatísticos e de aprendizado profundo aplicados à predição de casos de dengue. Contudo, sua abordagem baseia-se na transformação dos dados em uma matriz de correlação cruzada, afastando-se da modelagem temporal direta, além de empregar arquiteturas mais complexas que não são foco do presente estudo.

Diante desse panorama, conforme evidenciado no Quadro 1, verifica-se uma lacuna na literatura no que se refere à comparação direta entre um modelo estatístico clássico, como o ARIMA, e um modelo baseado em redes neurais recorrentes, como o LSTM, aplicados à predição de casos de dengue por meio de séries temporais. Nesse sentido, o presente trabalho diferencia-se ao integrar essas duas abordagens sob um mesmo conjunto de dados e métricas de avaliação, contribuindo para uma análise mais objetiva do desempenho dessas técnicas no contexto epidemiológico da dengue.

Quadro 1 – Análise e comparação dos trabalhos relacionados

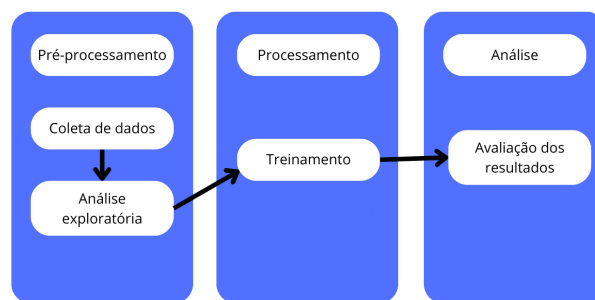
Trabalho	Área da Saúde	Série Temporal	Uso do modelo LSTM	Uso do modelo ARIMA	Predição de Dengue
Carvalho e Cortes (2025)	X	X	X	X	
Araújo <i>et al.</i> (2023)	X	X			
Silveira e Moreira (2020)	X				X
Zanardo <i>et al.</i> (2024)	X		X	X	X
Este trabalho	X	X	X	X	X

Fonte: Autoria própria (2025).

4 METODOLOGIA

Neste capítulo é apresentada a metodologia adotada neste estudo, estruturada em três etapas principais, conforme ilustrado na Figura 4: pré-processamento, processamento e análise. Cada etapa compreende um conjunto de procedimentos essenciais para a construção, treinamento e avaliação dos modelos preditivos ARIMA e LSTM aplicados a série temporal de notificações de casos de dengue no município de Mossoró.

Figura 4 – Etapas da metodologia



Fonte: Autoria própria (2025).

4.1 Pré-processamento

A primeira fase da etapa de pré-processamento iniciou-se com a coleta de dados, que corresponde à obtenção das informações utilizadas nesta pesquisa. Em seguida, foi realizada a análise exploratória, responsável pela transformação dos dados brutos em uma série temporal.

4.1.1 Coleta de dados

Os dados utilizados neste trabalho foram obtidos a partir de um projeto de pesquisa desenvolvido na Universidade Federal Rural do Semi-Árido (UFERSA). O conjunto de dados contempla notificações de casos de dengue no município de Mossoró, no estado do Rio Grande do Norte, constituindo uma base real de informações epidemiológicas e climáticas.

A base de dados abrange o período de janeiro de 2010 a dezembro de 2020 e está estruturada em formato tabular, composta por linhas e colunas. No total, a base contém 133 registros mensais, cada um representando um período específico, e oito variáveis associadas

às condições climáticas e epidemiológicas do município. As colunas correspondem ao ano da ocorrência, mês, índice de precipitação pluviométrica, umidade relativa do ar, temperatura média, número de casos notificados de dengue, número de casos acumulados nos 30 dias anteriores e número de casos acumulados nos 60 dias anteriores.

A Tabela 1 apresenta um recorte da base de dados, ilustrando os registros referentes ao ano de 2010, com o objetivo de exemplificar a estrutura e o tipo de informação disponível para análise.

Tabela 1 – Dados referentes ao ano de 2010 da base utilizada neste estudo

Ano	Mês	Chuvas	Umidade	Temp. Média	Casos	Casos 30d	Casos 60d
2010	jan	187.6	70.1828	27.40	6	5	5
2010	fev	32.8	69.8527	28.72	5	5	3
2010	mar	56.2	70.2903	28.87	5	3	9
2010	abr	249.9	76.3583	28.05	3	9	8
2010	mai	8.9	76.8804	27.43	9	8	25
2010	jun	51.2	69.6292	27.14	8	25	28
2010	jul	1.5	65.6277	26.92	25	28	18
2010	ago	0	60.2661	26.41	28	18	17
2010	set	0	59.1042	27.13	18	17	32
2010	out	52.0	66.2030	27.57	17	32	53
2010	nov	0	64.9931	27.88	32	53	194
2010	dez	19.4	69.1492	27.73	53	194	349

Diante das características do conjunto de dados, optou-se por sua utilização na construção de séries temporais, de modo a viabilizar o treinamento, a aplicação e a posterior avaliação do desempenho dos modelos ARIMA e LSTM com base nos dados empregados no processo de treinamento.

4.1.2 Análise exploratória

Nesta etapa, realizou-se uma análise exploratória preliminar dos dados brutos, bem como o pré-processamento necessário para sua transformação em séries temporais. Para esse fim, utilizou-se a linguagem de programação *Python*¹ e suas bibliotecas auxiliares *Pandas*², *Matplotlib*³ e *NumPy*⁴, empregadas tanto nesta etapa quanto nas demais fases da metodologia.

Inicialmente, com o auxílio da biblioteca *Pandas*, a base de dados foi carregada e

¹ Foi utilizada a versão 3.13.7 da linguagem Python, disponível em: <https://www.python.org/>.

² Utilizou-se a versão 2.3.3 da biblioteca *Pandas*, disponível em: <https://pandas.pydata.org/>.

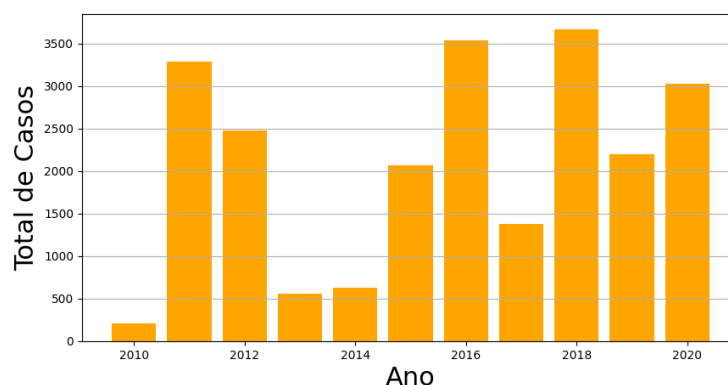
³ Foi empregada a versão 3.10.7 da biblioteca *Matplotlib*, disponível em: <https://matplotlib.org/>.

⁴ A versão da biblioteca *NumPy* utilizada neste trabalho foi a 2.3.4, disponível em: <https://numpy.org/pt/>.

submetida a uma agregação anual, na qual se realizou a soma do número de notificações de casos de dengue registrados em cada ano, obtendo-se assim o total anual de notificações.

A Figura 5 apresenta o gráfico de barras correspondente ao total anual de notificações de dengue no município de Mossoró, no período de 2010 a 2020. Observa-se que o ano de 2018 apresentou o maior número de registros, totalizando 3.665 casos, seguido pelos anos de 2016, com 3.541 casos, e 2011, com 3.290 casos.

Figura 5 – Total de notificações anuais de dengue entre os anos de 2010 e 2020



Fonte: Autoria própria (2025).

De modo geral, não se observa um padrão regular na variação anual do número de casos. Como exemplo, nota-se que após um elevado número de ocorrências em 2016, houve uma redução significativa em 2017, seguida de um novo aumento em 2018. Comportamento semelhante ocorre entre os anos de 2018, 2019 e 2020, enquanto em períodos anteriores essa dinâmica não se repete. Além disso, os picos de incidência ocorrem em intervalos irregulares, indicando a ausência de um comportamento cíclico claramente definido nos dados anuais.

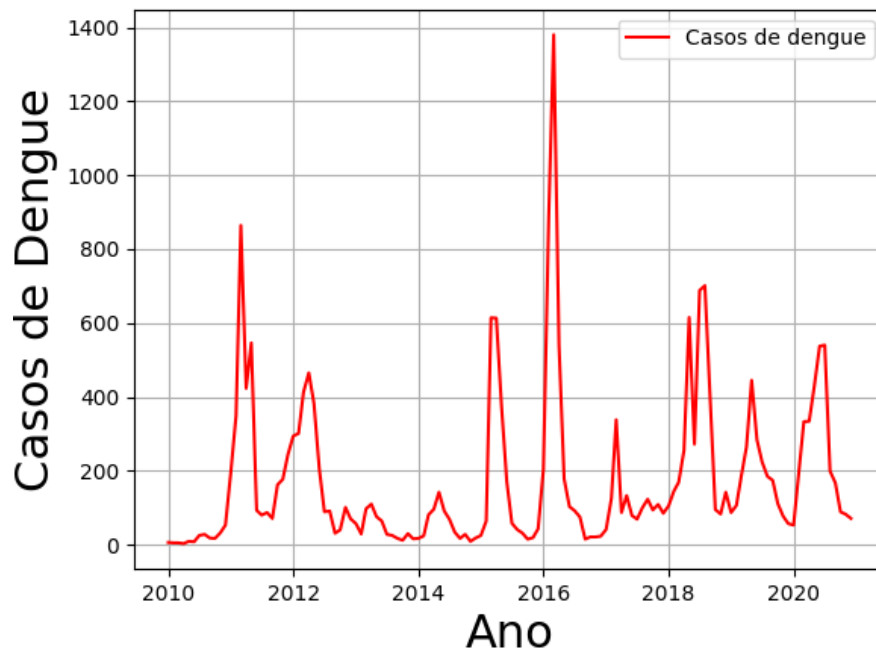
Após essa análise inicial, procedeu-se à transformação dos dados brutos em séries temporais. Para isso, a base foi organizada em uma estrutura do tipo *dataframe*, possibilitando a ordenação cronológica dos registros e a associação de valores mensais às respectivas datas.

Inicialmente, foi realizada a padronização da identificação dos meses, uma vez que estes se encontravam originalmente representados por abreviações textuais, como “jan”, “fev” e “mar”. Cada abreviação foi convertida para seu respectivo valor numérico, permitindo a combinação consistente das informações de ano e mês. Em seguida, essas informações foram transformadas em objetos do tipo *data*, adotando-se o primeiro dia de cada mês como referência, o que garantiu a padronização temporal dos registros. A coluna resultante foi então definida como índice da base de dados.

Posteriormente, foram removidas as colunas que não seriam utilizadas na modelagem, tais como as identificações originais de ano e mês e as variáveis auxiliares referentes aos casos acumulados em 30 e 60 dias anteriores, além das variáveis meteorológicas. Dessa forma, a base final manteve apenas as informações essenciais para a construção das séries temporais, compostas pelos valores mensais de notificações de casos de dengue entre os anos de 2010 a 2020.

Por fim, com o uso da biblioteca Matplotlib, foram geradas as séries temporais correspondentes às variáveis analisadas, conforme ilustrado na Figura 6.

Figura 6 – Séries temporais das notificações de dengue no período de 2010 a 2020 no município de Mossoró



Fonte: Autoria própria (2025).

4.2 Processamento

A etapa de processamento concentrou-se na fase de treinamento, na qual foram definidas e aplicadas as estratégias e modelagens necessárias para o treinamento de ambos os modelos.

4.2.1 Treinamento

Nesta etapa do processamento, foram definidas as abordagens de treinamento para ambos os modelos, bem como realizados os respectivos procedimentos de ajuste. Considerando as particularidades de cada abordagem, o modelo LSTM demandou um pré-processamento adicional, a fim de adequar os dados à escala apropriada para a rede neural e possibilitar a construção de sua arquitetura.

A mudança de escala é uma etapa fundamental em modelos baseados em redes neurais, pois esses modelos são sensíveis às magnitudes dos valores de entrada. Ao normalizar os dados, reduz-se o impacto de variações extremas, evita-se instabilidade numérica durante os cálculos internos e favorece uma convergência mais eficiente do algoritmo de treinamento, contribuindo para a melhoria do desempenho e da estabilidade do modelo. Por esse motivo, foi aplicado para este trabalho a mudança de escala na fase de treinamento do LSTM.

Todo o processo de pré-processamento e treinamento do modelo LSTM foi realizado com o auxílio das bibliotecas *Keras*¹ e *scikit-learn*², ambas pertencentes ao ecossistema de bibliotecas da linguagem Python, juntamente com o *NumPy* e o *Pandas*, já empregados em etapas anteriores da metodologia.

Por outro lado, para o modelo ARIMA, aplicou-se exclusivamente a técnica de *Grid Search*, com o objetivo de identificar os valores mais adequados para os parâmetros p , d e q . O parâmetro p refere-se à ordem do componente autorregressivo do modelo, indicando quantos valores passados da própria série temporal são utilizados para explicar o valor atual. O parâmetro d corresponde ao grau de diferenciação aplicado à série, sendo responsável por torná-la estacionária, ou seja, tornar a série constantes em relação a suas variações. Enquanto o parâmetro q representa a ordem do componente de média móvel, que considera a influência dos erros residuais de períodos anteriores.

A técnica de *Grid Search* consiste em uma busca exaustiva por combinações pré-definidas desses parâmetros, na qual o modelo é ajustado repetidamente utilizando diferentes valores de p , d e q . Para cada combinação testada, avalia-se o desempenho do modelo por meio de uma métrica de erro, permitindo selecionar a configuração que apresenta o melhor desempenho preditivo para a série temporal analisada.

O treinamento desse modelo foi realizado com o uso da biblioteca *statsmodels*, também

¹ Foi utilizada a versão 3.11.3 da biblioteca Keras. A documentação referente à camada LSTM encontra-se disponível em: https://keras.io/api/layers/recurrent_layers/lstm/.

² Utilizou-se a versão 1.7.2 da biblioteca scikit-learn, disponível em: <https://scikit-learn.org/>.

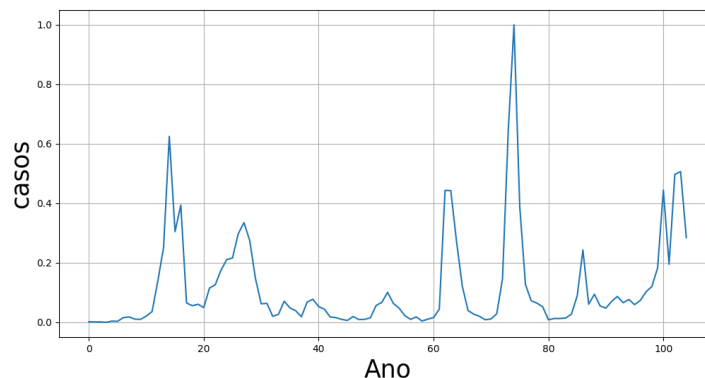
integrada ao ambiente Python. Ademais, para ambos os modelos, definiu-se uma divisão da série temporal em conjuntos de treinamento e teste. O conjunto de treinamento correspondeu aos 80% iniciais da série temporal das notificações de dengue, enquanto os 20% finais, compostos pelas observações mais recentes, foram reservados para a avaliação do desempenho dos modelos.

4.2.2 Treinamento do modelo LSTM

O treinamento do modelo LSTM teve início após a separação dos dados em conjuntos de treino e teste. Inicialmente, foi realizada a normalização da série temporal, de modo que os valores passassem a estar contidos no intervalo entre 0 e 1. Para isso, utilizou-se a função *MinMaxScaler*, disponibilizada pela biblioteca *scikit-learn*. Esse procedimento preserva a estrutura temporal da série, ao mesmo tempo em que ajusta a escala dos dados, conforme ilustrado na Figura 7.

Considerando que as notificações de dengue apresentam variações expressivas ao longo do tempo, a aplicação dessa técnica torna-se necessária para evitar instabilidades numéricas durante o treinamento e para melhorar a eficiência do processo de convergência da rede.

Figura 7 – Série temporal normalizada dos casos de dengue



Fonte: Autoria própria (2025).

Após o pré-processamento, foi aplicada a técnica de janela deslizante (*sliding window*) à série temporal normalizada. Essa abordagem consiste em utilizar um conjunto fixo de observações passadas como entrada do modelo, com o objetivo de prever o valor subsequente da série.

No presente trabalho, definiu-se uma janela de tamanho seis, o que implica que cada entrada do modelo é composta pelos valores correspondentes aos seis meses anteriores, enquanto o mês subsequente é utilizado como alvo da predição. Inicialmente, avaliou-se uma janela

de tamanho doze, equivalente a um ano completo; contudo, os resultados obtidos com essa configuração apresentaram desempenho inferior quando comparados aos da janela de tamanho seis. As métricas de avaliação indicaram que a utilização de um período muito longo dificultou a identificação de padrões locais na dinâmica dos casos, tornando a janela menor mais adequada para capturar variações de curto e médio prazo.

Com a definição das janelas, os dados de entrada foram reorganizados com o auxílio da biblioteca *NumPy*, de modo a atender ao formato tridimensional exigido pelo modelo LSTM, composto por amostras, passos de tempo e atributos. As amostras correspondem ao total de janelas geradas a partir da série temporal; os passos de tempo indicam a quantidade de observações passadas consideradas em cada janela, neste caso, seis; por fim, o número de atributos refere-se às variáveis disponíveis em cada passo de tempo. Neste estudo, foi considerado apenas um atributo, correspondente ao número de casos de dengue, não sendo realizado o cruzamento com outras variáveis, como precipitação ou umidade.

Em seguida, iniciou-se a construção e o treinamento da rede neural. O modelo adotado seguiu uma arquitetura sequencial, composta por quatro camadas LSTM empilhadas, cada uma contendo 50 unidades, responsáveis pelo aprendizado dos padrões temporais da série. Após cada camada recorrente, foi inserida uma camada *Dropout*, com taxa de desativação de 20%, com o objetivo de reduzir o risco de *overfitting* e evitar que o modelo se ajuste excessivamente aos dados de treinamento.

A escolha da quantidade de 50 unidades em cada camada foi definida a partir de testes computacionais preliminares realizados de forma empírica, nos quais se avaliou o desempenho do modelo frente a diferentes configurações arquiteturais. Esses testes permitiram identificar um equilíbrio adequado entre capacidade de aprendizado dos padrões temporais da série.

Após as camadas recorrentes, adicionou-se uma camada densa com um único neurônio, responsável pela geração das previsões. O modelo foi compilado utilizando o otimizador *Adam* e a função de perda MAE. O treinamento foi realizado ao longo de 150 épocas, com uma validação interna correspondente a 10% dos dados de treinamento.

Ao final do processo de treinamento, os valores previstos pelo modelo foram transformados de volta para a escala original, permitindo a comparação direta com os dados reais do conjunto de teste.

4.2.3 Treinamento do modelo ARIMA

O modelo ARIMA, diferentemente do modelo LSTM, não demandou etapas adicionais de pré-processamento, uma vez que a série temporal das notificações de dengue, sintetizada nas etapas anteriores, já se encontrava no formato exigido pela biblioteca *statsmodels* para a realização do ajuste e da predição.

Assim, o treinamento do modelo ARIMA concentrou-se na identificação dos valores mais adequados para seus parâmetros. Para esse fim, empregou-se a técnica de *Grid Search*, cuja lógica de funcionamento para este trabalho é apresentada no pseudocódigo ilustrado na Figura 10, disponível no Apêndice A.

Para a aplicação do *Grid Search* nesse trabalho, foram definidos intervalos de valores entre 0 e 5 para os parâmetros p , d e q . Esses intervalos foram estabelecidos com base em testes empíricos preliminares, nos quais se observou que valores superiores a 5 não resultavam em melhorias significativas no desempenho do modelo, além de acarretarem um aumento expressivo no custo computacional.

A partir desses intervalos, todas as combinações possíveis de parâmetros foram avaliadas, sendo que, a cada iteração, um modelo ARIMA era ajustado aos dados de treinamento. Em seguida, eram geradas previsões para o conjunto de teste e calculado o erro por meio da métrica MAE, de forma análoga ao procedimento adotado no treinamento do modelo LSTM. Sempre que o valor de MAE obtido se mostrava inferior ao menor erro registrado até então, a combinação correspondente de parâmetros era armazenada como a melhor configuração encontrada.

Como resultado desse processo, os melhores parâmetros identificados foram p igual a 2, d igual a 0 e q igual a 4. A escolha de d igual a zero indica que a série apresentou estacionariedade suficiente, dispensando a aplicação de diferenciação adicional. O componente autorregressivo de ordem dois permite que o modelo utilize os dois valores anteriores da série para explicar o valor atual, enquanto o componente de média móvel de ordem quatro incorpora a influência dos quatro últimos erros residuais, contribuindo para a correção de oscilações recentes não capturadas pela parte autorregressiva.

Por fim, com a obtenção dos melhores parâmetros por meio da aplicação do *Grid Search*, o modelo ARIMA foi ajustado utilizando a base de dados destinada ao treinamento, ou seja, os 80% iniciais da série, adotando a configuração *ARIMA*(2,0,4). Em seguida, foi realizada a comparação direta entre as previsões do modelo e os dados reais do conjunto de teste.

5 RESULTADOS

Este capítulo apresenta os resultados obtidos a partir dos experimentos realizados com os modelos LSTM e ARIMA, considerando exclusivamente o desempenho observado durante a fase de treinamento e validação. A análise concentra-se nas previsões geradas a partir da divisão da série temporal em conjuntos de treino e teste, bem como na comparação entre os valores previstos pelos modelos e os valores reais observados no período de validação. Além disso, são apresentadas métricas quantitativas de avaliação com o objetivo de mensurar a eficiência e a capacidade de ajuste de cada abordagem.

Diferentemente de uma predição futura, o objetivo desta etapa consiste em avaliar o quão bem cada modelo é capaz de aprender os padrões temporais presentes nos dados históricos e reproduzi-los no conjunto de teste. Dessa forma, busca-se realizar uma análise comparativa de desempenho, destacando as diferenças entre um modelo estatístico clássico e um modelo baseado em redes neurais recorrentes.

5.1 Avaliação dos resultados

A avaliação do desempenho dos modelos foi realizada por meio das métricas Raiz do Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE) e Erro Percentual Absoluto Médio (MAPE), amplamente utilizadas em problemas de previsão de séries temporais. Para o cálculo dessas métricas, foram empregadas funções disponibilizadas pelos pacotes pertencentes à biblioteca Keras.

Inicialmente, os valores reais do conjunto de teste e as previsões geradas por cada modelo durante a fase de treinamento foram utilizados como entrada para o cálculo das métricas. Os resultados obtidos encontram-se na Tabela 2.

Modelo	RMSE	MAE	MAPE
LSTM	77,61	77,61	27,83
ARIMA	134,26	95,26	54,31

Tabela 2 – Resultados das métricas RMSE, MAE e MAPE obtidos na fase de treinamento dos modelos LSTM e ARIMA.

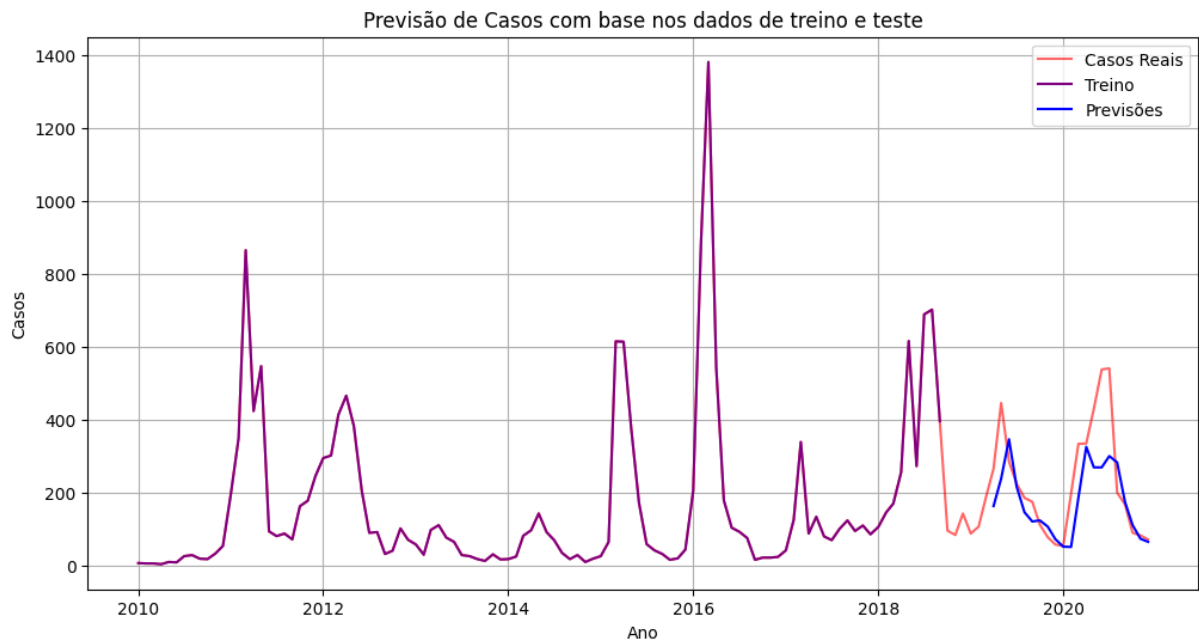
Com base nos valores apresentados na Tabela 2, observa-se que o modelo LSTM apresentou desempenho superior em todas as métricas avaliadas. O valor de RMSE obtido pelo modelo LSTM foi de 77,61, indicando uma menor dispersão entre os valores previstos e os valores reais,

o que reflete maior precisão nas previsões. Da mesma forma, o MAE, também com valor de 77,61, evidencia que, em média, o erro absoluto cometido pelo modelo foi relativamente baixo ao longo do conjunto de teste. O MAPE, com valor de 27,83%, reforça esse desempenho ao indicar um erro percentual médio significativamente menor, demonstrando maior aderência das previsões do LSTM aos dados reais.

Em contrapartida, o modelo ARIMA apresentou valores mais elevados em todas as métricas analisadas, com RMSE igual a 134,26, MAE de 95,26 e MAPE de 54,31%. Esses resultados indicam maior dispersão e erro médio nas previsões, bem como um erro percentual consideravelmente superior quando comparado ao modelo LSTM. Tal comportamento sugere uma menor capacidade do modelo ARIMA em capturar os padrões temporais presentes nos dados históricos, especialmente em períodos de maior variabilidade ou não linearidade da série.

De forma a complementar a análise quantitativa, a Figura 8 apresenta a comparação entre os dados de treino, os dados de teste e as previsões geradas pelo modelo LSTM durante a fase de treinamento.

Figura 8 – Comparativo dos resultados do treinamento do modelo LSTM, separados em treino, teste e predição



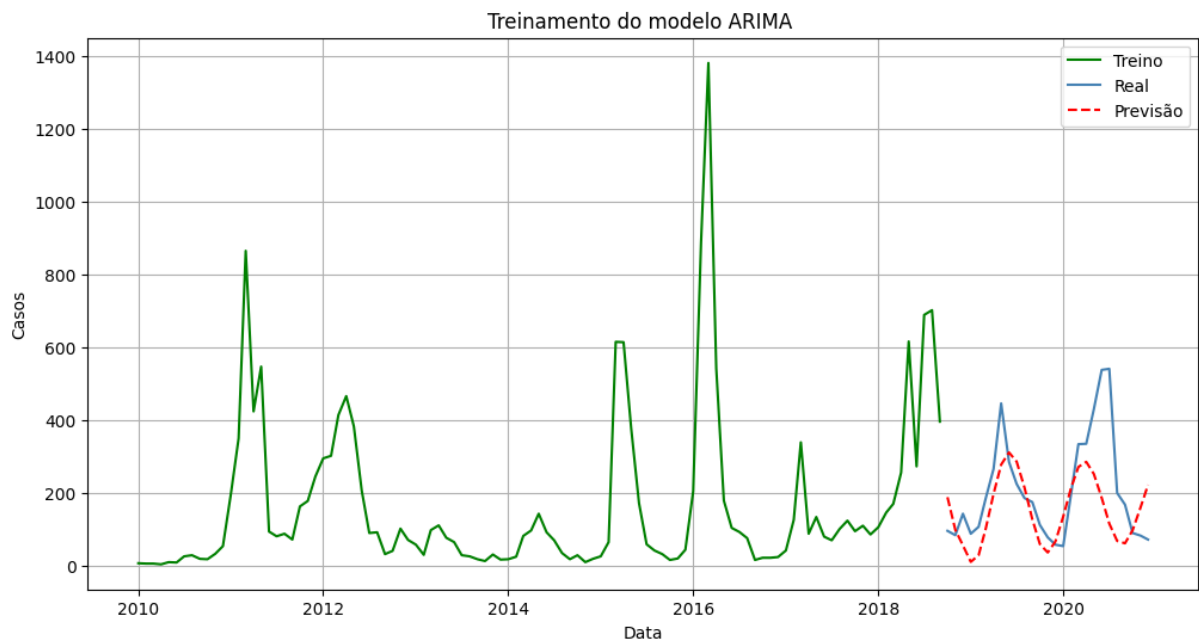
Fonte: Autoria própria (2025).

Convém ressaltar que a predição do modelo LSTM tem início a partir do ponto em que ocorre a separação entre os conjuntos de treinamento e teste. Entretanto, conforme observado na Figura 8, nota-se a presença de uma lacuna temporal antes do início efetivo das previsões.

Essa lacuna está diretamente associada à utilização da técnica de janela deslizante, previamente descrita, a qual foi definida com tamanho igual a 6. Dessa forma, o modelo necessita das seis observações imediatamente anteriores para compor a primeira entrada válida e, somente após esse intervalo, torna-se possível realizar a predição do valor subsequente.

No que se refere ao modelo ARIMA, a Figura 9 apresenta o comparativo entre os dados de treino, os dados de teste e as previsões geradas pelo modelo durante a fase de validação. Observa-se que, embora o modelo consiga capturar parcialmente a tendência da série, suas previsões apresentam maior afastamento em relação aos valores reais, especialmente em períodos de maior variação dos dados.

Figura 9 – Comparativo dos resultados do treinamento do modelo ARIMA, separados em treino, teste e predição



Fonte: Autoria própria (2025).

6 CONCLUSÕES E TRABALHOS FUTUROS

No que se refere aos objetivos propostos neste trabalho, estes foram plenamente alcançados. Inicialmente, realizou-se a obtenção da base de dados contendo o número de notificações reais de dengue. Em seguida, conforme previsto, os dados passaram por etapas de análise exploratória e preparação, de modo a viabilizar sua correta utilização. Após o tratamento, aplicaram-se técnicas de programação em Python, juntamente com bibliotecas auxiliares, com o intuito de estruturar os registros na forma de séries temporais.

Com a série temporal devidamente configurada, procedeu-se à implementação e ao treinamento dos modelos LSTM e ARIMA, permitindo a avaliação do desempenho preditivo de cada abordagem. Por fim, realizou-se a comparação entre os modelos, possibilitando identificar diferenças de desempenho e estabelecer considerações relevantes sobre suas capacidades preditivas no contexto das notificações de dengue no município de Mossoró.

Com base nos resultados obtidos, conclui-se que o modelo LSTM apresentou maior eficiência em suas previsões quando comparado ao modelo ARIMA, especialmente ao se considerar as métricas quantitativas de avaliação. O modelo baseado em redes neurais recorrentes obteve desempenho superior em todas as métricas analisadas, evidenciando maior capacidade de aprendizado dos padrões temporais presentes nos dados históricos de notificações de dengue.

Destaca-se, em particular, o fato de os valores de RMSE e MAE do modelo LSTM terem sido idênticos, o que indica a ausência de erros extremamente discrepantes em relação aos dados reais utilizados como conjunto de teste. Esse comportamento sugere maior estabilidade nas previsões e menor sensibilidade a valores atípicos. Ademais, o valor de MAPE obtido, aproximadamente 27,83%, indica que o modelo apresentou um erro percentual médio relativamente baixo, errando, em média, cerca de 28% das notificações quando comparado aos valores reais do período de validação.

Sob uma perspectiva qualitativa, observou-se que os modelos apresentaram comportamentos distintos em suas séries temporais preditas, refletindo as diferenças inerentes às suas abordagens. O modelo LSTM demonstrou maior capacidade de acompanhar as variações e oscilações ao longo do tempo, ajustando-se de forma mais precisa às mudanças abruptas e aos picos da série. Em contrapartida, o modelo ARIMA produziu previsões mais suavizadas, concentrando-se principalmente na identificação das tendências gerais da série, em vez de buscar a aproximação exata dos valores observados. Tal característica o torna adequado para análises de comportamento médio, porém limitado em cenários com elevada variabilidade temporal.

Dessa forma, os resultados indicam que, para o conjunto de dados analisado e no contexto do treinamento supervisionado, o modelo LSTM mostrou-se mais apropriado para a tarefa de predição de notificações de dengue, enquanto o modelo ARIMA pode ser considerado uma alternativa viável quando o objetivo principal estiver associado à análise de tendências temporais.

Convém ressaltar que, durante o desenvolvimento da pesquisa, identificou-se como limitação o uso exclusivo de séries temporais univariadas. Tal restrição impediu a incorporação de variáveis externas potencialmente relevantes, como índices pluviométricos, temperaturas médias e níveis de umidade do município de Mossoró. Essa limitação representa uma oportunidade significativa para aprofundamentos em trabalhos futuros.

Nesse sentido, como perspectivas de continuidade, sugere-se a ampliação do estudo para a utilização de séries temporais multivariadas, incorporando variáveis climáticas e ambientais, bem como a avaliação de outros modelos baseados em redes neurais, com o objetivo de comparar seus desempenhos com o modelo LSTM. Além disso, a expansão da base de dados histórica pode contribuir para um treinamento mais robusto, possibilitando a identificação de padrões temporais que não foram capturados no escopo desta pesquisa.

REFERÊNCIAS

- ANDRADE, P.; BRILHANTE, L. V. T. Predizendo epidemias de dengue, no distrito federal, utilizando algoritmos de regressão. **Monografia de Graduação, Universidade de Brasília**, v. 5, 2018.
- ARAÚJO, I. de L. *et al.* Comparação da performance de algoritmos de aprendizado de máquina para análise preditiva de febre amarela no estado de minas gerais. Universidade Federal de Minas Gerais, 2023.
- BOX, G.; JENKINS, G. Analysis: Forecasting and control. **San francisco**, 1976.
- BRASIL. Ministério da Saúde. **Boletins Epidemiológicos — Edições de 2019**. 2019. Disponível em: <https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2019>.
- BROCKWELL, P. J.; DAVIS, R. A. **Introduction to time series and forecasting**. [S.l.]: Springer, 2002.
- CARVALHO, D. M. de; CORTES, O. A. C. Uso de redes neurais recorrentes para predição de doenças isquêmicas do coração usando séries temporais. In: SBC. **Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)**. [S.l.], 2025. p. 91–96.
- CATARINO, M. H. **Redes Neurais**. [S.l.]: Freitas Bastos, 2025.
- CORREIA, J. B. *et al.* Comparing arima and lstm models to predict time series in the oil industry. In: SBC. **Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)**. [S.l.], 2021. p. 129–136.
- DONALISIO, M. R.; FREITAS, A. R. R.; ZUBEN, A. P. B. V. Arboviroses emergentes no brasil: desafios para a clínica e implicações para a saúde pública. **Revista de saúde pública**, SciELO Public Health, v. 51, p. 30, 2017.
- MELO, B. *et al.* Tendências epidemiológicas das arboviroses no brasil: Análise de dados de dengue, zika e chikungunya entre 2020 e 2022: Epidemiological trends of arboviruses in brazil: Analysis of data on dengue, zika and chikungunya between 2020 and 2022. **Revista de Ensino e Saúde na Amazônia**, v. 3, n. 1, 2025.
- NILSSON, N. J. **Principles of artificial intelligence**. [S.l.]: Morgan Kaufmann, 2014.
- PAIXÃO, G. M. de M. *et al.* Machine learning na medicina: Revisão e aplicabilidade. **Arquivos brasileiros de cardiologia**, SciELO Brasil, v. 118, n. 1, p. 95, 2022.
- RIBEIRO, N. F. Previsão de velocidade do vento com dados reais de um parque eólico localizado no ceará utilizando redes neurais artificiais: um estudo comparativo das redes mlp, lstm, gru e cnn. 2024.
- SILVA, I. N. d.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas**. [S.l.]: Artliber Editora, 2010.
- SILVEIRA, F. R. de V.; MOREIRA, L. Y. M. R. Utilização de algoritmos de aprendizagem de máquina na predição de arboviroses transmitidas pelo aedes aegypti. **Conexões-Ciência e Tecnologia**, v. 14, n. 1, p. 64–71, 2020.

SOUZA, C. C. de *et al.* Box-jenkins (arima) e redes neurais artificiais multilayer perceptron (rna_mlp) na previsão de preços de ações da bovespa. **UNICIÊNCIAS**, v. 28, n. 1, p. 06–14, 2024.

SOUZA, C. F. C. de *et al.* Previsão da incidência de câncer por gênero em belo horizonte: inovação no uso de redes neurais artificiais para planejamento de saúde pública. **ARACÊ**, v. 7, n. 3, p. 15184–15201, 2025.

SOUZA, R. C.; CAMARGO, M. E. Análise e previsão de séries temporais: os modelos arima. **Ijuí: Sedigraf**, 1996.

VIEIRA, G. F. F.; CASTRO, A. F. de; OLIVEIRA, A. G. de. Análise de dados de dengue, no município de mossoró, nordeste do brasil, entre os anos de 2015 e 2020, utilizando linguagem python. **Revista Caribeña de Ciencias Sociales**, v. 12, n. 6, p. 2513–2531, 2023.

ZANARDO, G. E. *et al.* Uma comparação entre métodos baseados em aprendizado de máquina para inferir número de casos semanais de dengue. In: SBC. **Seminário Integrado de Software e Hardware (SEMISH)**. [S.l.], 2024. p. 37–48.

APÊNDICE A – EXEMPLO DE CÓDIGO

Figura 10 – Pseudo-código da lógica do grid search para o treinamento do modelo ARIMA

```
# Listas de valores para testar no Grid Search
valores_para_p = [0, 1, 2, 3, 4, 5]
valores_para_d = [0, 1, 2, 3, 4, 5]
valores_para_q = [0, 1, 2, 3, 4, 5]

melhor_mae = infinito
melhor_parametro = nulo

para p faça valores_para_p:
  para d faça valores_para_d:
    para q faça valores_para_q:
      # Criar o modelo ARIMA com os parâmetros atuais
      modelo = ARIMA(treino, ordem = (p, d, q))

      # Ajustar o modelo aos dados de treino
      ajustar_modelo = modelo.ajustar()

      # Gerar previsões para o período correspondente ao conjunto de teste
      predicao = ajustar_modelo.prever(passos = tamanho(teste))

      # Calcular o erro MAE entre o valor real e o previsto
      mae_atual = MAE(teste, predicao)

      # Verificar se este modelo é melhor que o melhor encontrado até agora
      se mae_atual < melhor_mae:
        melhor_mae = mae_atual
        melhor_parametro = (p, d, q)
```

Fonte: Autoria própria.