**INSTITUTO FEDERAL DE ALAGOAS**

***CAMPUS* ARAPIRACA**

**CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**SAMILA RAPHAELA DE OLIVEIRA**

**VICTOR LUAN DE LIMA LEMOS**

# Aplicação de Modelagem Computacional e Inteligência Artificial na Classificação do Risco de Surtos de Doenças Infecciosas na Cidade de Maceió, Alagoas

**ARAPIRACA, AL**

**2026**

SAMILA RAPHAELA DE OLIVEIRA
VICTOR LUAN DE LIMA LEMOS

# APLICAÇÃO DE MODELAGEM COMPUTACIONAL E INTELIGÊNCIA ARTIFICIAL NA CLASSIFICAÇÃO DO RISCO DE SURTOS DE DOENÇAS INFECCIOSAS NA CIDADE DE MACEIÓ, ALAGOAS

Trabalho de Conclusão de Curso apresentado ao Curso Superior de Sistemas de Informações do Instituto Federal de Alagoas, Campus Arapiraca, como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Társis Marinho de Souza

Coorientadora: Profª. Drª. Cledja Karina Rolim da Silva

ARAPIRACA, AL
2026

**Dados Internacionais de Catalogação na Publicação**
**Instituto Federal de Alagoas**
***Campus* Arapiraca**

004
O48a Oliveira, Samila Raphaela de.

Aplicação de modelagem computacional e inteligência artificial na classificação de risco de surtos de doenças infecciosas na cidade de Maceió, Alagoas / Samilla Raphaela de Oliveira, Victor Luan de Lima Lemos – Dados eletrônicos (1 arquivo : 3,1 MB). – 2026.

Sistema requerido: Adobe Acrobat Reader.
Modo de acesso: Internet.
Orientação: Prof. Dr. Tarsis Marinho de Souza.
Co-orientador: Profª. Drª.Cledja Karina Rolim da Silva.

Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Instituto Federal de Alagoas, *Campus* Arapiraca, Arapiraca, 2026.

1. Análise de dados. 2. Doenças infecciosas. 3. Visualização de dados. 4. Aprendizado de máquina. 5. Saúde. I. Lemos, Victor Luan de Lima. II. Título..

**Luciete Barbosa da Silva | Bibliotecária – CRB-4/1739**

SAMILA RAPHAELA DE OLIVEIRA
VICTOR LUAN DE LIMA LEMOS

# APLICAÇÃO DE MODELAGEM COMPUTACIONAL E INTELIGÊNCIA ARTIFICIAL NA CLASSIFICAÇÃO DO RISCO DE SURTOS DE DOENÇAS INFECCIOSAS NA CIDADE DE MACEIÓ, ALAGOAS

Trabalho de Conclusão de Curso apresentado ao Curso Superior de Sistemas de Informações do Instituto Federal de Alagoas, Campus Arapiraca, como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Társis Marinho de Souza

Coorientadora: Profª. Drª. Cledja Karina Rolim da Silva

Aprovado(a) em: 20/02/2026.

**BANCA EXAMINADORA**

Documento assinado digitalmente
gov.br
TARSIS MARINHO DE SOUZA
Data: 23/02/2026 14:35:28-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Társis Marinho de Souza (Orientador)
Instituto Federal de Alagoas - IFAL

Documento assinado digitalmente
gov.br
CLEDJA KARINA ROLIM DA SILVA
Data: 23/02/2026 09:25:24-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Cledja Karina Rolim da Silva (Coorientadora)
Instituto Federal de Alagoas - IFAL

Documento assinado digitalmente
gov.br
EDVONALDO HORACIO DOS SANTOS
Data: 23/02/2026 11:19:26-0300
Verifique em https://validar.iti.gov.br

Prof. Me. Edvonaldo Horário dos Santos
Instituto Federal de Alagoas - IFAL

Documento assinado digitalmente
gov.br
LEONARDO SOARES E SILVA
Data: 23/02/2026 19:24:30-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Leonardo Soares e Silva
Instituto Federal de Pernambuco - IFPE

## AGRADECIMENTOS

Agradeço primeiramente à minha família. Meu pai, Jorge Samuel (*in memoriam*), minha maior inspiração nos estudos e na vida, cuja presença permanece em cada conquista. Juntamente, agradeço à mulher mais forte que conheço, minha mãe, Cida, que, ao lado dele, preparou o terreno para que eu pudesse trilhar meus próprios passos. Ao meu irmão, Jorge, por servir de referência e por caminhar comigo nessa jornada.

Agradeço também aos meus colegas de faculdade, que dividiram comigo não apenas os trabalhos e provas, mas também as experiências que marcaram essa etapa da minha vida. Em especial, à minha dupla nos trabalhos e na vida, Victor, pela parceria ao longo desses anos. Aos nossos amigos, Eduardo e Guilherme, por serem apoio constante e por transformarem os desafios em momentos mais leves e memoráveis.

Agradeço também ao meu orientador, professor e amigo, Társis, que ministrou minha primeira aula de Lógica de Programação ainda no ensino médio, plantando a semente que me conduziu até este curso.

Estendo minha gratidão ao nosso laboratório, o DIT, espaço onde encontrei pessoas incríveis e dividimos momentos inesquecíveis, que ampliaram minha visão e fortaleceram minha caminhada.

Também agradeço à nossa coorientadora, Cledja, cuja trajetória inspira tantas pessoas e, em especial, as mulheres do curso. Que cada vez mais deixemos de ser uma modesta minoria e passemos a ocupar com ainda mais protagonismo os espaços na área de TI.

Por fim, agradeço ao IFAL Campus Arapiraca, que há uma década integra minha história e foi essencial para a construção da minha trajetória acadêmica e profissional. A todos os amigos, colegas e professores que fizeram parte desse percurso, minha profunda gratidão.

*Samila Raphaela de Oliveira*

## AGRADECIMENTOS

Agradeço ao professor Tarsis, meu orientador e amigo, por ter enxergado em mim, ainda em 2020, um potencial que eu mesmo não conseguia reconhecer. Sua confiança foi determinante para os caminhos que percorri desde então, gerando frutos que ultrapassam a dimensão acadêmica e impactam profundamente minha vida pessoal e profissional. Seu olhar atento, humano e acolhedor transforma trajetórias todos os dias e eu sou prova disso.

Agradeço também à professora Cledja, cuja disciplina, dedicação e desenvoltura em sala de aula sempre foram fonte de inspiração. Sua postura ética e comprometida, dentro e fora da universidade, revela a essência de quem possui verdadeira vocação para educar.

Ambos representam, com excelência, o que significa ter alma de educador.

À minha mãe, Gildete, e à minha irmã, Taisy, deixo minha gratidão mais sincera. Duas mulheres que sempre me ensinaram, pelo exemplo, que o caminho do estudo é transformador. Foram apoio constante, incentivo diário e base sólida para que eu pudesse chegar até aqui. Ao meu avô, Pedro, que foi e sempre será uma figura paterna em minha vida, agradeço pelo cuidado, pela força e pela serenidade ensinada diante das falhas e frustrações.

À minha dupla, Samila, expresso meu orgulho por ti e alegria por compartilhar não apenas este trabalho, mas toda a trajetória da graduação. Sua parceria, dedicação e amizade tornaram essa caminhada mais significativa. Nada disso teria sido possível sem você. Agradeço também a Eduardo e Guilherme, cujo companheirismo, risadas e leveza tornaram os desafios da graduação mais suportáveis e os momentos memoráveis.

Ainda, estendo minha gratidão a todos os amigos e colegas do DIT/IFAL. A convivência, os projetos desenvolvidos em conjunto e até mesmo os momentos simples de descontração contribuíram para minha formação acadêmica e profissional. Fazer parte de um núcleo tão comprometido e especial foi, sem dúvida, um dos maiores privilégios desta jornada.

*Victor Luan de Lima Lemos*

## RESUMO

Ao longo da história, as doenças infecciosas vêm representando uma ameaça significativa à população mundial, permanecendo como um desafio relevante e prioritário para a saúde pública, o que corrobora a necessidade de estratégias inovadoras voltadas à prevenção, à vigilância e ao controle. No Brasil, apesar da existência de sistemas nacionais de vigilância epidemiológica, a ocorrência de surtos e a reemergência de agravos infecciosos evidenciam limitações e fragilidades na capacidade de antecipação e de resposta diante de eventos epidêmicos. Nesse contexto, a identificação precoce de padrões de risco é de fundamental importância para subsidiar ações oportunas em saúde pública. Diante desse cenário, os modelos baseados em inteligência artificial destacam-se por sua capacidade de identificar padrões complexos e por fornecer subsídios ao planejamento de ações em saúde pública, apresentando desempenho comparável ou superior às abordagens estatísticas tradicionais e tornando-se ferramentas essenciais no monitoramento e no controle de epidemias. Sendo assim, este trabalho tem como objetivo analisar e comparar modelos computacionais e algoritmos de aprendizado de máquina aplicados à classificação do risco de surtos de doenças infecciosas, a partir de dados provenientes do Sistema de Informação de Agravos de Notificação (SINAN). Com isso, busca-se contribuir para o avanço da modelagem computacional em saúde pública e para o fortalecimento de estratégias baseadas em evidências no monitoramento epidemiológico.



**Palavras-chave:** doenças infecciosas; análise de dados; visualização de dados; saúde; aprendizado de máquina.

## ABSTRACT

Infectious diseases have historically represented a significant threat to populations worldwide and remain a relevant and priority challenge for public health, reinforcing the need for innovative strategies focused on prevention, surveillance, and control. In Brazil, despite the existence of national epidemiological surveillance systems, the occurrence of outbreaks and the reemergence of infectious diseases reveal limitations and vulnerabilities in the capacity to anticipate and respond to epidemic events. In this context, the early identification of risk patterns is essential to support timely public health actions. Within this scenario, models based on artificial intelligence have gained prominence due to their ability to identify complex patterns and support public health planning. These approaches have demonstrated performance comparable to or superior to traditional statistical methods, becoming essential tools for epidemic monitoring and control. Therefore, this study aims to analyze and compare computational models and machine learning algorithms applied to the classification of the risk of infectious disease outbreaks, using data from the Notifiable Diseases Information System (SINAN). The study seeks to contribute to the advancement of computational modeling in public health and to the strengthening of evidence-based strategies in epidemiological surveillance.



**Keywords:** infectious diseases; data analysis; data visualization; health; machine learning.

## LISTA DE FIGURAS

## LISTA DE QUADROS E TABELAS

**SUMÁRIO**

## 1 INTRODUÇÃO

Ao longo das últimas décadas, a persistência das doenças infecciosas como um dos principais problemas de saúde pública mundial tem demandado o desenvolvimento de estratégias cada vez mais sofisticadas para vigilância, prevenção e controle de doenças e de agravos. Estima-se que tais doenças sejam responsáveis por milhões de óbitos registrados anualmente, especialmente em regiões marcadas por desigualdades socioeconômicas e limitações estruturais nos sistemas de saúde (JONES *et al*., 2008).

Associado a esse cenário, o aumento da mobilidade populacional e a intensificação dos processos de globalização têm contribuído de forma significativa para a rápida disseminação de agentes infecciosos em diferentes escalas geográficas. Estudos demonstram que os fluxos de deslocamento humano influenciam diretamente a propagação espacial de epidemias, reduzindo a eficácia de estratégias baseadas exclusivamente em análises retrospectivas (BALCAN *et al*., 2009).

Nesse contexto, a capacidade de compreender e de antecipar padrões de transmissão torna-se um elemento central para a mitigação de surtos e de epidemias (SANTANGELO *et al*., 2023). Dessa forma, evidencia-se a necessidade de abordagens que possibilitem a antecipação de cenários epidemiológicos, apoiando ações preventivas mais eficazes (SANTANGELO *et al*., 2023).

Historicamente, a modelagem epidemiológica tem sido amplamente empregada para descrever e para analisar a dinâmica de transmissão de doenças infecciosas. Modelos matemáticos compartimentais, como o SIR (*Susceptible, Infected, Recovered*) e o SEIR (*Susceptible, Exposed, Infected, Recovered*), estruturam a população em compartimentos que representam estágios distintos da dinâmica de transmissão. No modelo SIR, os indivíduos transitam entre os estados de suscetível, infectado e recuperado, enquanto o modelo SEIR acrescenta o compartimento de expostos, representando o período de incubação antes do indivíduo tornar-se infeccioso. Embora esses modelos ofereçam forte fundamentação teórica e interpretabilidade epidemiológica, apresentam limitações

quando aplicados a cenários altamente complexos, nos quais múltiplos fatores interagem de forma não linear e variam ao longo do tempo.

No contexto brasileiro, a disponibilidade de sistemas nacionais de vigilância epidemiológica, como o Sistema de Informação de Agravos de Notificação (SINAN), possibilita a exploração de séries históricas extensas e detalhadas. Entretanto, a heterogeneidade regional, as variações na qualidade dos dados e as particularidades locais exigem abordagens adaptadas às realidades específicas de cada estado. Em regiões como Maceió, Alagoas, a aplicação de técnicas computacionais avançadas podem auxiliar na compreensão da dinâmica de transmissão e no fortalecimento da vigilância epidemiológica regional.

Aliado a isso, a crescente diversidade de técnicas computacionais disponíveis para previsão epidemiológica impõe o desafio de selecionar métodos que sejam não apenas precisos, mas também robustos e adequados às características dos dados analisados. Diferentes algoritmos podem apresentar comportamentos distintos frente a variações na qualidade dos dados, na presença de ruído e na dependência temporal das séries epidemiológicas. Assim, a comparação sistemática entre abordagens de modelagem, aliada à análise de diferentes estratégias de treinamento e validação, torna-se essencial para compreender os limites e potencialidades de cada método, evitando generalizações indevidas e contribuindo para a adoção de soluções mais confiáveis em contextos reais de vigilância em saúde (HASTIE, 2009; SHAMAN; KARSPECK, 2012).

Nesse cenário, a avaliação comparativa entre diferentes técnicas de modelagem torna-se um aspecto fundamental não apenas para fins acadêmicos, mas também para a definição de estratégias preditivas adequadas a contextos epidemiológicos específicos, como o analisado neste estudo. A literatura aponta que não existe um modelo universalmente superior, sendo o desempenho fortemente influenciado pela natureza dos dados disponíveis, pela granularidade temporal e espacial considerada e pelas estratégias de treinamento adotadas (SHAMAN; KARSPECK, 2012; ZHANG *et al*., 2020 apud SANTANGELO *et al*., 2023). No entanto, observa-se que muitos estudos concentram-se na aplicação isolada de modelos ou na avaliação baseada exclusivamente em métricas globais, sem incorporar critérios adicionais de confiabilidade e interpretação operacional.

Diante dessa lacuna, torna-se relevante investigar, de forma sistemática e comparativa, diferentes abordagens de modelagem aplicadas ao contexto local, incorporando não apenas métricas de desempenho, mas também a análise de incerteza das previsões, de modo a subsidiar decisões mais robustas em vigilância epidemiológica.

## 1.1 OBJETIVO GERAL

Este trabalho tem como objetivo prospectar soluções baseadas em modelos preditivos para doenças infecciosas no município de Maceió, Alagoas, por meio da aplicação e comparação de técnicas de modelagem computacional e aprendizado de máquina, visando estimar a evolução temporal da transmissão e analisar o desempenho de diferentes abordagens no contexto da vigilância em saúde pública. Busca-se, portanto, investigar e comparar distintas tecnologias e estratégias de modelagem, identificando aquelas que apresentam melhor desempenho preditivo e maior adequação às características dos dados epidemiológicos disponíveis, contribuindo para o fortalecimento da tomada de decisão baseada em evidências.

## 1.2 OBJETIVOS ESPECÍFICOS

Para este projeto, foram definidos os seguintes objetivos específicos:

- Prospectar modelos computacionais para previsão de doenças infecciosas, incluindo modelos epidemiológicos clássicos, algoritmos de aprendizado de máquina supervisionado e abordagens baseadas em séries temporais.
- Comparar diferentes abordagens de modelagem por meio de métricas apropriadas de avaliação preditiva, considerando aspectos como erro, capacidade de generalização e estabilidade dos modelos.
- Investigar o impacto de diferentes estratégias de treinamento, validação e parametrização sobre o desempenho dos modelos, identificando configurações mais adequadas aos dados epidemiológicos analisados.

# 2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados as principais concepções e perspectivas relacionadas aos temas aqui abordados. Os conceitos apresentados a seguir

fornecem o embasamento teórico necessário para a aplicação da análise de dados e para o desenvolvimento da solução proposta, abrangendo os fundamentos da modelagem computacional e os aspectos relacionados à aprendizagem de máquina.

## 2.1 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é um campo da Computação que busca compreender e construir entidades inteligentes, isto é, sistemas capazes de realizar tarefas associadas a capacidades cognitivas como percepção, aprendizagem, raciocínio e tomada de decisão. Além de ser uma área ampla, a IA abrange diversos subcampos e se torna relevante para diferentes esferas da atividade intelectual humana, ao sistematizar e automatizar processos intelectuais (GOMES, 2010).

Uma forma consolidada de estruturar teoricamente a IA é a perspectiva de agentes inteligentes, na qual um agente é definido como uma entidade que recebe percepções do ambiente e executa ações, e a IA é apresentada como o estudo e a engenharia desses agentes. Nessa perspectiva, a avaliação de um agente depende de elementos como medida de desempenho, conhecimento disponível sobre o ambiente e sensores e atuadores componentes que tornam explícitas as restrições e incertezas envolvidas no problema (RUSSELL; NORVIG, 2004 apud GOMES, 2010).

### 2.1.1 Aprendizagem de Máquina

O Aprendizado de Máquina (AM) constitui uma das abordagens centrais da Inteligência Artificial contemporânea, pois viabiliza que computadores adquiram habilidades a partir de exemplos (dados), em vez de depender exclusivamente de regras explícitas codificadas por programadores (LUDERMIR, 2021). Do ponto de vista conceitual, o AM pode ser compreendido como um subcampo da computação que busca integrar técnicas matemáticas e estatísticas com algoritmos computacionais para identificar padrões e produzir modelos úteis para classificação, predição ou detecção. Em aplicações práticas, a lógica do AM contrasta com sistemas baseados em regras, pois o conhecimento necessário à tarefa é inferido do conjunto de dados disponível, ao invés de ser inteiramente especificado a priori. Assim, o AM se torna particularmente pertinente quando há grande quantidade de

dados e a complexidade do fenômeno dificulta a formalização completa em regras determinísticas (PAIXÃO *et al*., 2022).

Um ponto crítico associado à avaliação é o equilíbrio entre ajuste aos dados de treinamento e capacidade de generalização, pois certos modelos supervisionados podem apresentar tendência ao sobreajuste (overfitting), quando aprendem padrões específicos (inclusive ruído) do treino e perdem desempenho em dados novos. Nesse cenário, a etapa de avaliação com dados de teste (separados do treino) é essencial para estimar a robustez do modelo e reduzir a chance de conclusões indevidas sobre sua qualidade (PAIXÃO *et al*., 2022).

#### 2.1.1.1 Tipos de aprendizado

A literatura recente descreve que os métodos de Aprendizado de Máquina podem ser classificados conforme a disponibilidade de rótulos e o tipo de feedback fornecido ao algoritmo durante o processo de aprendizagem. Essa tipologia é relevante porque orienta a escolha de técnicas, a definição do desenho experimental e os critérios de avaliação, já que diferentes problemas exigem estratégias distintas de aprendizagem, conforme discutido por Paixão *et al*. (2022).

Segundo Ludermir (2021) e Menezes *et al*. (2024), o aprendizado supervisionado ocorre quando o conjunto de treinamento contém pares entrada–saída, isto é, cada exemplo possui um rótulo (resposta desejada), e o objetivo do modelo é aprender uma função que generalize para novos dados. Nesse paradigma, é comum distinguir tarefas de classificação (saídas discretas) e regressão (saídas contínuas), como forma de caracterizar o tipo de variável alvo e o tipo de erro que será minimizado.

O aprendizado não supervisionado é caracterizado pela ausência de rótulos, de modo que o algoritmo busca identificar estruturas intrínsecas nos dados, como padrões, agrupamentos e representações de menor dimensionalidade. Esse tipo de aprendizado é frequentemente empregado para exploração de dados e para a descoberta de padrões latentes, exigindo interpretação posterior no contexto do domínio (LUDERMIR, 2021).

Já o aprendizado por reforço envolve um agente que aprende por interação com o ambiente, recebendo sinais de recompensa ou punição como forma de feedback para ajustar suas decisões ao longo do tempo. Nesse paradigma, a ênfase recai sobre a construção de uma política de ações que maximize

recompensas acumuladas, sendo comum sua aplicação em problemas como controle, robótica e jogos, onde as ações influenciam o estado do ambiente (LUDERMIR, 2021).

De acordo com Paixão *et al*. (2022) e Ludermir (2021), além desses três tipos clássicos, a literatura contemporânea também descreve aprendizado semissupervisionado, no qual o treinamento combina pequena quantidade de dados rotulados com grande volume de dados não rotulados, estratégia útil quando rotular exemplos é caro ou demorado. Assim, a escolha do tipo de aprendizado depende de fatores como disponibilidade e qualidade de rótulos, custo de anotação, natureza do problema e objetivo analítico, influenciando diretamente o desenho metodológico e a avaliação do modelo.

#### 2.1.1.2 Treinamento de modelos

De acordo com Ludermir (2021), o treinamento de modelos em Aprendizado de Máquina corresponde à etapa na qual um algoritmo ajusta seus parâmetros internos a partir de um conjunto de dados, com o objetivo de aprender padrões que representem adequadamente o problema em análise. Nesse processo, o modelo utiliza exemplos previamente disponíveis para estabelecer relações entre variáveis de entrada e saída, constituindo a base do aprendizado orientado por dados.

De forma geral, o treinamento envolve a apresentação iterativa dos dados ao algoritmo, permitindo que os parâmetros do modelo sejam progressivamente ajustados conforme os exemplos observados. Esse ajuste é guiado por uma função que expressa o grau de adequação do modelo aos dados, possibilitando que o aprendizado ocorra de maneira gradual e controlada (FONTANA, 2020).

A literatura destaca que o treinamento não se resume à execução automática do algoritmo, mas exige decisões metodológicas, como a definição do conjunto de dados utilizado, a configuração inicial do modelo e a escolha do procedimento de aprendizado. De acordo com Paixão *et al.* (2022), essas decisões influenciam diretamente o comportamento do modelo treinado e sua capacidade de representar o fenômeno estudado.

Além disso, o treinamento de modelos é compreendido como um processo iterativo, no qual ajustes sucessivos são realizados até que o modelo atinja

um desempenho considerado satisfatório segundo critérios previamente definidos para o problema proposto. Fontana (2020) ressalta que essa característica reforça a importância do treinamento como etapa central no desenvolvimento de soluções baseadas em Aprendizado de Máquina.

Assim, o treinamento de modelos constitui o elo fundamental entre os dados disponíveis e a construção efetiva de modelos computacionais capazes de aprender a partir desses dados, servindo de base para as etapas posteriores de avaliação e validação, tratadas nos próximos subtópicos (LUDERMIR, 2021).

#### 2.1.1.3 *Overfitting*, *underfitting* e generalização

No contexto do Aprendizado de Máquina, a generalização refere-se à capacidade de um modelo treinado de manter um bom desempenho quando aplicado a dados não vistos durante o treinamento (objetivo fundamental de modelos úteis). Essa característica é essencial, pois modelos que apenas memorizam os exemplos do conjunto de treinamento, sem capturar os padrões subjacentes, têm desempenho limitado em casos reais (PAIXÃO *et al*., 2022).

O fenômeno de overfitting, também chamado de sobreajuste, ocorre quando um modelo aprende de forma excessiva as peculiaridades específicas do conjunto de treinamento, incluindo ruídos ou variações aleatórias, de modo que sua performance em dados não vistos se deteriora. Esse comportamento é típico de modelos muito complexos, com grande número de parâmetros, que acabam “decorando” os dados em vez de inferir padrões gerais, comprometendo sua capacidade de generalização. (LUDERMIR, 2021; SCHLEDER; FAZZIO, 2021)

Por outro lado, o underfitting, ou subajuste, acontece quando um modelo é incapaz de aprender os padrões relevantes dos dados de treinamento, resultando em baixo desempenho tanto no treinamento quanto em dados novos. Tal condição usualmente se manifesta quando a estrutura do modelo é simples demais ou quando não existe informação suficiente para suportar a identificação das relações presentes nos dados (LUDERMIR, 2021; SCHLEDER; FAZZIO, 2021).

De acordo com Ludermir (2021), a relação entre overfitting e underfitting está ligada ao equilíbrio entre a capacidade do modelo e a complexidade dos dados: modelos muito simples podem subestimar os padrões existentes, enquanto modelos excessivamente complexos podem ajustar-se demais aos dados de treino, falhando em prever novos exemplos com precisão.

Portanto, o objetivo durante o desenvolvimento e avaliação de modelos de Aprendizado de Máquina é encontrar um “ajuste” intermediário que maximize a capacidade de generalização, ou seja, a habilidade de o modelo representar corretamente as relações subjacentes nos dados e aplicá-las de forma robusta em situações inéditas (PAIXÃO *et al.*, 2022)

2.1.1.4 Validação (hold-out, k-fold, validação temporal)

Segundo Ludermir (2021), no Aprendizado de Máquina, a validação corresponde ao conjunto de procedimentos utilizados para estimar o desempenho de um modelo em dados não utilizados durante o treinamento, sendo uma etapa fundamental para avaliar sua capacidade de generalização. Esses métodos permitem analisar se o modelo aprendido representa adequadamente o fenômeno estudado ou se seu resultado obtido está restrito aos dados utilizados no ajuste dos parâmetros.

Um dos métodos mais simples de validação é o hold-out, no qual o conjunto de dados é dividido em partes distintas, geralmente destinadas ao treinamento e à avaliação do modelo. Nesse procedimento, o modelo é treinado com uma parcela dos dados e avaliado com o restante, fornecendo uma estimativa direta de capacidade de generalização. Apesar de sua simplicidade, o método hold-out pode ser sensível à forma como os dados são particionados, especialmente em bases pequenas (CERRI, CARVALHO, 2017).

De acordo com Cerri & Carvalho (2017), a validação cruzada k-fold é uma estratégia mais robusta, na qual o conjunto de dados é dividido em k subconjuntos, sendo que, a cada iteração, um subconjunto é utilizado para validação enquanto os demais são usados para treinamento. Esse processo é repetido k vezes, permitindo que todos os dados sejam utilizados tanto para treinamento quanto para validação em diferentes momentos. A média dos resultados obtidos fornece uma estimativa mais estável da qualidade preditiva do modelo.

Em problemas que envolvem dados temporais, a literatura ressalta a inadequação de métodos de validação que realizam embaralhamento aleatório dos dados, pois isso pode violar a ordem cronológica das observações. Nesses casos, emprega-se a validação temporal, na qual o modelo é treinado com dados de períodos anteriores e avaliado com dados de períodos posteriores, preservando a estrutura temporal do fenômeno analisado (LUDERMIR, 2021).

Assim, a escolha do método de validação deve considerar a natureza dos dados, o tamanho da base e o tipo de problema em estudo, pois diferentes estratégias podem produzir estimativas distintas de desempenho. A aplicação adequada de técnicas de validação é essencial para garantir a confiabilidade dos resultados obtidos com modelos de Aprendizado de Máquina (PAIXÃO *et al*., 2022).

#### 2.1.1.5 Métricas de avaliação

Paixão (2022) descreve que as métricas de avaliação em Aprendizado de Máquina são utilizadas para quantificar o desempenho de um modelo, permitindo analisar o quão adequadamente ele representa o fenômeno estudado e apoia a comparação entre diferentes modelos ou configurações. Essas métricas fornecem critérios objetivos para interpretar os resultados obtidos durante a fase de avaliação, sendo fundamentais para a análise da qualidade das previsões realizadas pelo modelo.

No contexto de tarefas de classificação, a literatura destaca métricas como acurácia, precisão, revocação (recall) e medida F1, cada uma enfatizando aspectos distintos do desempenho do modelo. A acurácia expressa a proporção de previsões corretas em relação ao total de exemplos avaliados, enquanto precisão e revocação permitem analisar, respectivamente, a confiabilidade das predições positivas e a capacidade do modelo em identificar corretamente os casos relevantes (CERRI; CARVALHO, 2017).

Segundo Ludermir (2021), a medida F1 combina precisão e revocação em um único valor, sendo especialmente útil em cenários nos quais há desbalanceamento entre classes, situação comum em aplicações reais de Aprendizado de Máquina. Em problemas multiclasse, a literatura também descreve diferentes formas de agregação da medida F1, como F1-macro, F1-micro e F1 ponderado. O F1-macro corresponde à média simples do F1 calculado para cada classe, atribuindo o mesmo peso a todas elas, sendo indicado quando se deseja avaliar o desempenho global independentemente da frequência das classes. O F1-micro agrega os totais de verdadeiros positivos, falsos positivos e falsos negativos antes do cálculo, tornando-se mais sensível às classes majoritárias. Já o F1 ponderado considera o suporte de cada classe no cálculo da média, equilibrando a influência das classes conforme sua representatividade no conjunto de dados (MENEZES *et al*., 2024).

Em problemas de regressão, são comumente empregadas métricas como erro médio absoluto (MAE) e erro quadrático médio (MSE), que medem a diferença entre os valores previstos pelo modelo e os valores reais observados. O MSE penaliza erros maiores de forma mais intensa, devido à elevação ao quadrado das diferenças. Derivado dessa métrica, o erro quadrático médio da raiz (RMSE) corresponde à raiz quadrada do MSE, apresentando a vantagem de manter a mesma unidade da variável predita, o que facilita sua interpretação prática (PAIXÃO *et al*., 2022).

Outra métrica amplamente utilizada é o coeficiente de determinação ($R^2$), que expressa a proporção da variabilidade dos dados explicada pelo modelo, permitindo avaliar seu poder explicativo em relação à média observada. Valores mais próximos de 1 indicam maior capacidade de explicação da variância dos dados (CERRI; CARVALHO, 2017).

Em contextos de previsão, utiliza-se também o erro percentual absoluto médio (MAPE), que mede o erro médio em termos percentuais em relação aos valores reais observados. Essa métrica é particularmente útil para comparar desempenho entre séries com diferentes escalas, embora apresente limitações quando os valores reais assumem valores muito baixos (MENEZES *et al*., 2024).

Assim, as métricas de avaliação desempenham papel central na análise dos resultados de modelos de Aprendizado de Máquina, pois fornecem subsídios quantitativos para a interpretação do desempenho obtido. A seleção criteriosa das métricas deve estar alinhada ao tipo de problema, aos objetivos do estudo e às características do conjunto de dados, contribuindo para conclusões mais consistentes e fundamentadas (CERRI; CARVALHO, 2017).

#### 2.1.1.6 Modelos de classificação

Os modelos de classificação constituem uma das principais abordagens do aprendizado de máquina supervisionado, sendo aplicados quando o objetivo é atribuir instâncias a categorias previamente definidas com base em variáveis explicativas. Conforme apresentado por Cerri e Carvalho (2017), a classificação representa uma das tarefas centrais do aprendizado de máquina, amplamente utilizada para identificação de padrões discriminativos em conjuntos de dados rotulados.

Entre os métodos tradicionais destaca-se a Regressão Logística, modelo probabilístico que estima a probabilidade de ocorrência de um evento por meio da função logística. Esse modelo estabelece uma relação linear entre as variáveis explicativas e o logaritmo da razão de chances, favorecendo a interpretabilidade e a análise da influência individual dos preditores (LUDERMIR, 2021). Contudo, sua estrutura linear pode limitar a modelagem de interações complexas e relações não lineares.

No contexto dos métodos baseados em árvores, o Random Forest caracteriza-se como um método de combinação de classificadores, construído a partir de múltiplas árvores de decisão treinadas com reamostragem dos dados e seleção aleatória de variáveis, buscando reduzir a variância e aumentar a capacidade de generalização do modelo (BREIMAN, 2001). Essa abordagem apresenta desempenho consistente em problemas com padrões não lineares e atributos heterogêneos.

De forma complementar, o Gradient Boosting também integra a família dos métodos de combinação de modelos, adotando uma estratégia sequencial na qual novos classificadores são ajustados para corrigir os erros do modelo anterior, resultando em elevada capacidade preditiva (FRIEDMAN, 2001).

Assim, a escolha entre Regressão Logística, Random Forest e Gradient Boosting envolve o equilíbrio entre interpretabilidade, complexidade do modelo e capacidade de generalização, aspectos relevantes em aplicações voltadas à análise de dados em saúde.

#### 2.1.1.7 Detecção de anomalias

A detecção de anomalias, também denominada detecção de *outliers*, constitui uma abordagem não supervisionada do aprendizado de máquina voltada à identificação de observações que apresentam comportamento significativamente distinto em relação ao padrão predominante dos dados. Segundo Chandola, Banerjee e Kumar (2009), anomalias podem indicar eventos raros, erros de medição ou mudanças estruturais no comportamento do sistema analisado, sendo amplamente estudadas em diferentes áreas aplicadas.

Entre os métodos modernos de detecção de anomalias destaca-se o *Isolation Forest*, proposto por Liu, Ting e Zhou (2008). Diferentemente de técnicas baseadas em densidade ou distância, esse algoritmo parte do princípio de que

observações anômalas são mais suscetíveis a isolamento por meio de partições aleatórias. O método constrói múltiplas árvores de isolamento, nas quais os dados são recursivamente particionados; instâncias que requerem menor número de divisões para serem isoladas tendem a ser classificadas como anômalas. Essa estratégia apresenta vantagem computacional em bases de maior dimensão e não depende explicitamente de estimativas de densidade.

Outra abordagem amplamente utilizada é o *Local Outlier Factor* (LOF), proposto por Breunig *et al*. (2000). O LOF fundamenta-se na comparação da densidade local de uma observação com a densidade de seus vizinhos mais próximos. Uma instância é considerada anômala quando apresenta densidade significativamente inferior à de seu entorno, caracterizando-se como um outlier local. Diferentemente de métodos globais, o LOF é capaz de identificar anomalias em regiões específicas do espaço de atributos, sendo particularmente útil em conjuntos de dados com distribuição heterogênea.

#### 2.1.1.8 Balanceamento de Dados com SMOTE

Em problemas de aprendizagem de máquina, a falta de dados causa o desbalanceamento de classes, fenômeno que ocorre quando uma ou mais classes apresentam número significativamente inferior de instâncias em relação às demais. Conforme discutido por Wang *et al.* (2021), essa situação pode comprometer o desempenho de algoritmos tradicionais de aprendizado de máquina, uma vez que muitos métodos de treinamento tendem a favorecer a classe majoritária durante o processo de otimização, resultando em baixa capacidade de identificação da classe minoritária.

Entre as técnicas desenvolvidas para lidar com esse problema, destaca-se o SMOTE (*Synthetic Minority Over-sampling Technique*). Diferentemente das abordagens baseadas na simples replicação de amostras da classe minoritária, o SMOTE realiza a geração de novas instâncias sintéticas por meio de interpolação entre exemplos existentes dessa classe.

O funcionamento do algoritmo consiste, inicialmente, na identificação dos $k$ vizinhos mais próximos de cada instância pertencente à classe minoritária. Em seguida, seleciona-se aleatoriamente um desses vizinhos e gera-se uma nova

amostra sintética ao longo do segmento de reta que conecta os dois pontos no espaço de atributos. A nova instância $P_{ij}$ pode ser descrita pela Equação (1):

$$p_{ij} = x_i + \lambda (x_{ij} - x_i) \quad (1)$$

Onde $x_i$ representa uma instância da classe minoritária, $x_{ij}$ corresponde a um de seus vizinhos mais próximos e $\lambda$ é um valor aleatório no intervalo [0,1].

Esse procedimento promove a expansão da classe minoritária de forma mais distribuída no espaço de características, aumentando sua representatividade sem duplicação direta de registros. Como discutido por Wang *et al*. (2021), a técnica busca preservar as características estruturais da classe ao mesmo tempo em que amplia sua densidade amostral. Todavia, o autor também expõe limitações do SMOTE tradicional, entre elas a possibilidade de geração de amostras sintéticas próximas às fronteiras entre classes, o que pode afetar a separabilidade dos dados. Além disso, o desempenho do método pode ser influenciado pela escolha do número de vizinhos $k$, parâmetro que impacta diretamente a dispersão das novas instâncias geradas.

2.1.1.9 Otimização e validação de hiperparâmetros

O desempenho de modelos de aprendizado de máquina depende não apenas dos dados utilizados para treinamento, mas também da escolha adequada de seus hiperparâmetros. Diferentemente dos parâmetros internos aprendidos automaticamente durante o treinamento, os hiperparâmetros são definidos previamente pelo pesquisador e influenciam diretamente o comportamento do algoritmo, como a taxa de aprendizado, a profundidade de árvores ou o número de vizinhos em métodos baseados em distância.

A seleção inadequada desses valores pode resultar em modelos subajustados (underfitting) ou sobreajustados (overfitting). Assim, técnicas sistemáticas de busca são empregadas para identificar combinações que maximizem o desempenho preditivo.

Nesse contexto, destaca-se o GridSearchCV, uma estratégia de otimização que realiza a avaliação exaustiva de diferentes combinações de hiperparâmetros previamente definidas em uma grade de busca. Para cada combinação possível, o modelo é treinado e avaliado, permitindo a seleção daquela

que apresenta melhor desempenho segundo uma métrica estabelecida (AHMAD *et al*, 2022).

Para garantir maior robustez na avaliação, o processo de busca é geralmente associado à validação cruzada (*cross-validation*). Para Ahmad *et al*. (2022), a combinação entre GridSearchCV e validação cruzada constitui uma estratégia estruturada para otimização de hiperparâmetros, permitindo identificar configurações que equilibram desempenho preditivo e capacidade de generalização.

Entretanto, a etapa de busca depende diretamente do esquema de validação adotado. Em problemas envolvendo séries temporais, a validação cruzada tradicional baseada em particionamento aleatório pode introduzir vazamento de informação, pois permite que dados futuros influenciem o treinamento. Para contornar essa limitação, utiliza-se o *TimeSeriesSplit*, método que preserva a ordem cronológica dos dados ao dividir o conjunto em blocos sequenciais. Em cada iteração, o modelo é treinado com observações passadas e validado em períodos posteriores, respeitando a estrutura temporal da série (SCURSONE *et al*., 2025).

Dessa forma, a combinação entre *GridSearch* e validação temporal constitui uma estratégia metodológica consistente para ajuste de hiperparâmetros em contextos nos quais a dependência temporal dos dados não pode ser ignorada.

## 2.2 DATASUS

O Departamento de Informática do Sistema Único de Saúde (DATASUS) constitui a principal infraestrutura de informação em saúde do Brasil, sendo responsável por concentrar e disponibilizar dados provenientes de diferentes sistemas nacionais utilizados no âmbito do Sistema Único de Saúde (SUS). Essas bases de dados são amplamente empregadas em estudos epidemiológicos, análises estatísticas e pesquisas em saúde pública, configurando-se como uma das principais fontes secundárias de informação em saúde no país (LIMA *et al*., 2021).

De acordo com Lima *et al*. (2021), o uso das bases disponibilizadas pelo DATASUS possibilita análises abrangentes sobre a distribuição de agravos, morbidade e mortalidade, além de permitir avaliações temporais e espaciais de eventos em saúde. Os autores destacam que a amplitude e a padronização desses dados favorecem sua aplicação em pesquisas científicas, especialmente em estudos de vigilância em saúde e planejamento de políticas públicas.

Estudos recentes publicados em periódicos científicos nacionais também evidenciam a relevância do uso de dados secundários oriundos de sistemas de informação em saúde, como os integrados ao DATASUS, para a produção de conhecimento científico. Pesquisas como a de Aguiar *et al.* (2022) demonstram que essas bases são frequentemente utilizadas em análises populacionais e estudos aplicados, reforçando seu papel como suporte para investigações em saúde coletiva e gestão em saúde.

Além disso, trabalhos publicados em periódicos da área da saúde, como os Arquivos Brasileiros de Cardiologia, indicam que bases nacionais de dados em saúde são amplamente empregadas em estudos clínicos e epidemiológicos, contribuindo para a análise de fatores associados a doenças e para a avaliação de desfechos em populações específicas. Esses estudos reforçam a importância de sistemas nacionais de informação como fonte para pesquisas baseadas em dados observacionais (KOIKE, 2025).

Apesar de suas potencialidades, a literatura ressalta que o uso de dados provenientes do DATASUS exige cuidados metodológicos, uma vez que se trata de informações oriundas de registros administrativos e de vigilância, sujeitos a limitações como subnotificação, inconsistências de preenchimento e variações na qualidade dos dados ao longo do tempo. Assim, análises baseadas nessas bases devem considerar tais restrições para garantir interpretações adequadas e resultados confiáveis (LIMA *et al*., 2021).

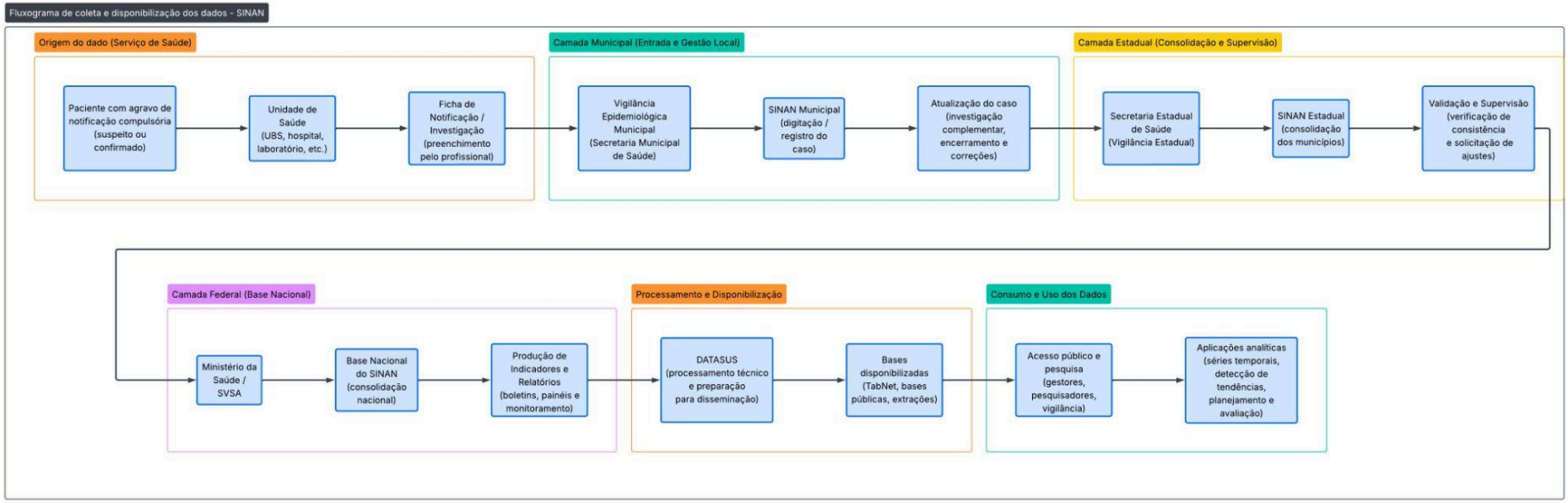
### 2.2.1 Sistema de Informação de Agravos de Notificação

O Sistema de Informação de Agravos de Notificação (SINAN) é um sistema nacional de informação em saúde que registra casos de agravos e doenças de notificação compulsória no Brasil, organizado para fornecer dados epidemiológicos que apoiem a vigilância em saúde pública. O SINAN foi concebido como uma ferramenta para coletar, transmitir e consolidar informações a partir de fichas de notificação preenchidas por profissionais de saúde, com o objetivo de produzir indicadores que contribuam para o conhecimento do perfil de agravos e para o planejamento de ações de controle e prevenção (ROCHA; BARTHOLOMAY *et al*., 2020; SILVA, 2016).

De acordo com Maia *et al*. (2014), a notificação compulsória de agravos no SINAN segue uma rotina padronizada em todo o território nacional, permitindo

comparações temporais e geográficas na ocorrência de eventos de saúde. Esse fluxo de informações envolve a alimentação regular das bases de dados por unidades de saúde municipais, com subsequente consolidação em esferas estadual e federal, o que favorece a análise da situação epidemiológica e a identificação de tendências, incluindo a detecção precoce de surtos. Estudos avaliativos também apontam que a implantação do SINAN enfrenta desafios relacionados à completude e à qualidade dos dados, sobretudo em localidades de menor estrutura de vigilância, sendo necessária constante avaliação e aperfeiçoamento.

Figura 1 – Fluxograma do SINAN




**Fonte:** Autores**.**

A utilização de bases como o SINAN tem sido amplamente explorada em pesquisas epidemiológicas no Brasil, especialmente em estudos que investigam a distribuição e determinantes de doenças como tuberculose e outras agravos de relevância pública, demonstrando sua importância como fonte de dados para análises de incidência, planejamento de políticas de saúde e avaliação de desempenho das ações de vigilância epidemiológica (BRITO *et al*., 2023; ROCHA; BARTHOLOMAY *et al*., 2020).

#### 2.2.1.1 Subnotificação

A subnotificação refere-se à ocorrência de casos de agravos ou doenças que, embora presentes na população, não são devidamente registrados nos sistemas oficiais de vigilância epidemiológica, como o Sistema de Informação de Agravos de Notificação (SINAN). Esse fenômeno compromete a capacidade do sistema em representar adequadamente a magnitude e a distribuição dos agravos, estando associado a falhas em diferentes etapas do processo de vigilância, como o acesso da população aos serviços de saúde, a detecção diagnóstica e o registro efetivo das informações nos sistemas de notificação (SILVA *et al*., 2020).

No contexto brasileiro, estudos apontam que a subnotificação constitui uma limitação relevante dos sistemas de vigilância epidemiológica, podendo afetar de forma transversal diferentes agravos de notificação compulsória. Essa limitação tornou-se mais evidente durante a pandemia de COVID-19, período em que se observaram alterações significativas no funcionamento dos serviços de saúde e das rotinas de vigilância. Análises baseadas em séries temporais demonstraram um decréscimo expressivo nas notificações compulsórias registradas no Brasil em 2020, quando comparadas aos padrões observados nos anos anteriores à pandemia (SALLAS, 2022).

Segundo Sallas (2022), as reduções observadas nas notificações durante o período pandêmico excederam as variações esperadas com base nas tendências epidemiológicas históricas, sugerindo que parte desse declínio pode ser atribuída à subnotificação, e não exclusivamente a uma redução real da ocorrência dos agravos. Esse fenômeno foi identificado de forma ampla, envolvendo diferentes grupos de doenças e agravos sob vigilância, o que reforça a interpretação de um impacto sistêmico da pandemia sobre a capacidade de notificação.

A literatura também destaca que a emergência sanitária da COVID-19 provocou efeitos indiretos relevantes sobre os programas de controle de doenças de notificação compulsória. De acordo com Borges (2024), a sobrecarga dos serviços de saúde, a reorganização das rotinas assistenciais e de vigilância e a priorização de recursos humanos e materiais para o enfrentamento da pandemia comprometeram a detecção e o registro oportuno de outros agravos. Além disso, a redução da procura da população pelos serviços de saúde durante esse período pode ter limitado a identificação de casos, contribuindo para o agravamento da subnotificação.

Dessa forma, embora o SINAN represente um instrumento central para o monitoramento epidemiológico e o planejamento de ações em saúde pública no Brasil, a subnotificação configura-se como uma limitação importante a ser considerada na interpretação dos dados. Evidências indicam que esse problema foi intensificado no contexto da pandemia de COVID-19, refletindo impactos estruturais e operacionais sobre os sistemas de vigilância epidemiológica. Assim, análises baseadas em dados do SINAN devem incorporar uma leitura crítica desse contexto, especialmente em estudos que envolvem comparações temporais ou avaliação de tendências epidemiológicas.

## 2.3 *DATA LAKE*

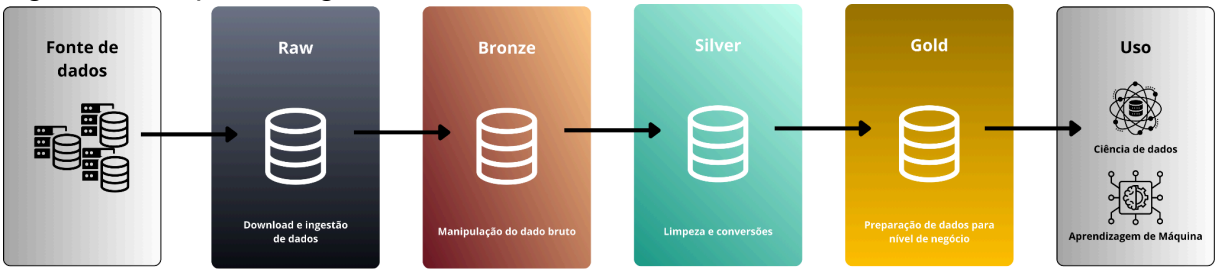
O conceito de Data Lake foi inicialmente apresentado por James Dixon, então diretor de tecnologia da Pentaho, em 2010, como uma alternativa às arquiteturas tradicionais de armazenamento de dados, caracterizada pela possibilidade de armazenar grandes volumes de dados em seu formato bruto, sem a necessidade de definição prévia de esquemas (DIXON, 2010). Essa abordagem visa preservar a granularidade e a diversidade dos dados, permitindo que diferentes usos analíticos sejam explorados posteriormente, conforme as necessidades do projeto.

Inmon (2016) amplia essa definição ao destacar que o Data Lake separa claramente o armazenamento dos dados de sua modelagem semântica, conferindo maior flexibilidade para análises exploratórias, ciência de dados e aprendizagem de máquina. Segundo o autor, essa separação reduz o acoplamento entre dados e aplicações analíticas, favorecendo a adaptação da arquitetura a novos requisitos e cenários de uso ao longo do tempo. Essa capacidade de integração é especialmente

relevante em domínios complexos, nos quais os dados são heterogêneos e evoluem continuamente, como ocorre em sistemas de informação em saúde.

Complementarmente, Nargesian *et al*. (2019) ressaltam que arquiteturas baseadas em Data Lake tendem a adotar zonas ou camadas de processamento, refletindo diferentes estágios de transformação dos dados. Essa organização contribui para a rastreabilidade, governança e reprodutibilidade das análises, aspectos essenciais em aplicações científicas e em soluções baseadas em aprendizagem de máquina.

Figura 2 – Arquitetura genérica de data lake




**Fonte:** Autores**.**

A Figura 2, representa o fluxo de processamento dos dados desde a fonte de dados até sua utilização em ciência de dados, aprendizado de máquina, entre outros. A camada *Raw* corresponde à etapa de ingestão dos dados brutos. A camada *Bronze* contempla a manipulação inicial e organização dos arquivos originais. A camada *Silver* envolve processos de limpeza, padronização e conversão dos dados. A camada *Gold* representa a preparação final dos dados, incluindo agregações e criação de atributos analíticos. Por fim, a etapa de *Uso* compreende as atividades de análise exploratória e aplicação de modelos de aprendizado de máquina.

Dessa forma, o Data Lake pode ser compreendido como um componente estruturante da arquitetura analítica moderna, que vai além do simples armazenamento de dados. Ao fornecer suporte à integração, organização e evolução dos dados ao longo de seu ciclo de vida, essa abordagem sustenta os processos de engenharia de dados e viabiliza a construção de modelos analíticos e preditivos robustos, como aqueles empregados em contextos de saúde pública.

### 2.3.1 Engenharia de Dados

Um pipeline de dados pode ser definido como um conjunto organizado de etapas responsáveis pela coleta, preparação, transformação, armazenamento e disponibilização de dados, de modo a garantir que informações brutas sejam convertidas em dados estruturados e aptos para análise e tomada de decisão (SANTOS; COSTA, 2020).

De acordo com Silva *et al*. (2021), pipelines de dados permitem automatizar fluxos de processamento, reduzir erros manuais e assegurar maior consistência no tratamento das informações. Ao estruturar o processamento em etapas bem definidas, torna-se possível monitorar a qualidade dos dados ao longo de todo o fluxo, além de facilitar a reprodutibilidade das análises realizadas.

De forma geral, a literatura descreve o pipeline de dados como composto por fases interdependentes. A primeira etapa é a ingestão de dados, na qual informações são coletadas a partir de diferentes fontes, como sistemas transacionais, bases públicas, arquivos estruturados ou dados abertos governamentais. Nessa fase, a confiabilidade das fontes e a integridade dos dados são aspectos críticos, pois influenciam diretamente as etapas posteriores do pipeline (SANTOS; COSTA, 2020).

Após a ingestão, ocorre a fase de processamento e transformação dos dados, que envolve atividades de limpeza, padronização, normalização e integração de diferentes conjuntos de dados. Essa etapa é essencial para eliminar inconsistências, tratar valores ausentes e adequar os dados aos objetivos analíticos do sistema, sendo destacada por Batista e Silva (2022) como determinante para a qualidade das análises realizadas.

Em seguida, os dados processados são direcionados para a etapa de armazenamento, geralmente em repositórios voltados à análise, como data warehouses, data lakes ou bases analíticas específicas. Esses ambientes são projetados para facilitar consultas, análises históricas e integração com ferramentas analíticas, assegurando desempenho e escalabilidade no acesso às informações (BATISTA; SILVA, 2022).

A etapa final do pipeline corresponde à disponibilização e consumo dos dados, na qual as informações tratadas passam a ser utilizadas por aplicações analíticas, sistemas de informação ou modelos de aprendizado de máquina. Nesse contexto, Silva *et al*. (2021) ressaltam que pipelines de dados desempenham papel

central ao fornecer dados consistentes, atualizados e confiáveis para análises avançadas e suporte à tomada de decisão.

Além do aspecto técnico, a literatura ressalta que pipelines de dados contribuem para a governança da informação, pois permitem rastrear a origem dos dados, documentar transformações realizadas e aumentar a transparência dos processos analíticos (BATISTA; SILVA, 2022).

Dessa forma, a adoção de pipelines de dados bem estruturados é considerada essencial para garantir a qualidade, a consistência e a confiabilidade das informações utilizadas em análises computacionais, apoiando tanto a produção científica quanto o desenvolvimento de sistemas inteligentes, conforme apontado por Silva *et al*. (2021).

#### 2.3.1.1 Feature Engineering

No contexto de *Data Lakes*, o *feature engineering* pode ser entendido como o conjunto de transformações que convertem dados brutos (frequentemente heterogêneos e em diferentes níveis de granularidade) em atributos (*features*) adequados para treinamento e inferência de modelos de *machine learning*. A arquitetura de *data lake* favorece esse processo ao armazenar dados em formatos abertos e com flexibilidade de *schema-on-read*, viabilizando exploração e transformação conforme a necessidade analítica; ao mesmo tempo, essa flexibilidade exige mecanismos de governança e camadas de processamento para tornar os dados confiáveis para uso em ML (ARMBRUST *et al*., 2021).

Na prática, o feature engineering em ambientes baseados em data lake tende a ocorrer em pipelines de processamento que derivam tabelas/visões mais “curadas” a partir da camada bruta, permitindo padronização, reprodutibilidade e rastreabilidade das transformações. Em uma visão de pipeline de reutilização de dados, a literatura descreve o *data lake* como componente que armazena dados e metadados em formato bruto e detalhado, enquanto camadas posteriores (por exemplo, datamarts e estruturas para features) apoiam a transformação dos dados em formas mais diretamente consumíveis por análises e modelos (LAMER *et al*., 2024).

Um desafio recorrente é garantir que as features usadas no treinamento sejam consistentes com as usadas em produção e que respeitem a correção

temporal (*point-in-time correctness*), principalmente quando há séries temporais e janelas de agregação. Trabalhos sobre features stores discutem esse problema como parte do gerenciamento de pipelines de ML, destacando a necessidade de padronizar a construção e o reuso de features para melhorar manutenção e reprodutibilidade dos modelos (ORR *et al*., 2021).

Além disso, pipelines de feature engineering frequentemente dependem de operações como *point-in-time join* para montar conjuntos de treino sem vazamento de informação do futuro (*data leakage*). Pesquisas em sistemas (PVLDB) tratam esses joins como operação crítica em *pipelines de feature stores* e mostram que otimizações específicas podem melhorar desempenho e apoiar a construção correta de conjuntos de treinamento em escala (LIU *et al*., 2023).

#### 2.3.1.2 Reprodutibilidade

No contexto de ambientes analíticos baseados em *Data Lake*, a reprodutibilidade refere-se à capacidade de obter os mesmos resultados analíticos a partir dos mesmos dados e procedimentos, mesmo quando executados em momentos distintos ou por diferentes pesquisadores. Em cenários de ciência orientada por dados e inteligência artificial, a reprodutibilidade depende da preservação e documentação dos artefatos computacionais que compõem o fluxo analítico, incluindo dados de entrada, código, parâmetros, versões de dependências e ambiente de execução, conforme discutido na literatura sobre e-Science (FERREIRA; VANZ, 2025).

A literatura destaca que, em infraestruturas de dados complexas, como *data lakes*, a simples descrição metodológica não é suficiente para garantir reprodutibilidade. É necessário que os processos de ingestão, transformação e disponibilização dos dados sejam padronizados, versionados e rastreáveis, de modo que a reconstrução de conjuntos analíticos possa ser realizada de forma consistente. Nesse sentido, práticas como versionamento de código, uso de ambientes computacionais controlados e registro sistemático de metadados são apontadas como elementos centrais para sustentar a reprodutibilidade computacional em pesquisas orientadas por dados (FERREIRA; VANZ, 2025).

Em aplicações na área da saúde, a organização estrutural dos fluxos de dados assume papel central para a consistência e o reúso analítico. A implementação de um *Data Lake* para dados em saúde descrita por Pagotto *et al*.

(2024) evidencia a importância de etapas claramente definidas de planejamento, ingestão, tratamento, armazenamento e disponibilização dos dados, com documentação dos processos e padronização dos fluxos. Embora o estudo não trate diretamente da reprodutibilidade computacional, ele sustenta empiricamente que a existência de rotinas organizadas e bem documentadas é condição necessária para que análises possam ser executadas novamente, auditadas e reutilizadas em contextos de decisão e pesquisa em saúde.

Além disso, discussões contemporâneas no campo da epidemiologia e da saúde coletiva reforçam que a incorporação de métodos computacionais e de inteligência artificial deve estar associada a compromissos explícitos com qualidade, transparência e rigor metodológico. Boing e Fonseca (2025) argumentam que o avanço dessas abordagens no contexto brasileiro exige práticas que permitam verificação e validação dos resultados produzidos, especialmente quando estes subsidiam decisões em saúde pública. Nesse sentido, a possibilidade de reproduzir análises a partir das mesmas bases de dados e procedimentos não constitui apenas um princípio epistemológico, mas um requisito prático para a auditoria de modelos analíticos e para a avaliação da confiabilidade dos resultados gerados.

Assim, no âmbito técnico de um *Data Lake*, a reprodutibilidade pode ser compreendida como um requisito transversal, sustentado por: (i) organização estruturada das camadas de dados; (ii) padronização e documentação dos processos de ingestão e transformação; (iii) versionamento de dados e código; e (iv) registro de informações sobre o ambiente computacional e os parâmetros de execução. Essas práticas contribuem para a confiabilidade, a auditabilidade e o uso sustentável de dados em projetos de modelagem computacional e inteligência artificial aplicados à área da saúde (FERREIRA; VANZ, 2025; PAGOTTO *et al*., 2024; BOING; FONSECA, 2025).

# 3 TRABALHOS RELACIONADOS

A previsão e o monitoramento de doenças infecciosas têm sido amplamente estudados na literatura científica, especialmente no contexto da vigilância epidemiológica e do apoio à tomada de decisão em saúde pública. As abordagens existentes variam quanto à formulação do problema, às técnicas de modelagem adotadas e aos critérios de avaliação utilizados. De maneira geral, os trabalhos podem ser agrupados em duas vertentes principais:

1. métodos que tratam a previsão como uma tarefa de regressão, buscando estimar quantitativamente o número de casos futuros; e
2. métodos que reformulam o problema como classificação ou detecção de risco, com o objetivo de identificar a ocorrência ou a iminência de surtos.

Considerando essa diversidade metodológica, este capítulo apresenta os trabalhos mais relevantes relacionados ao problema investigado, organizados segundo essas duas abordagens predominantes. Inicialmente, descreve-se o protocolo adotado para o levantamento e seleção dos estudos analisados. Em seguida, são discutidas as características, contribuições e limitações dos trabalhos baseados em regressão e classificação. Por fim, é realizada uma análise comparativa que fundamenta o posicionamento metodológico adotado neste estudo.

## 3.1 PROTOCOLO DE PESQUISA

O protocolo de pesquisa foi estruturado com o objetivo de orientar, de forma sistemática, as etapas de levantamento conceitual, definição metodológica e condução dos experimentos computacionais. Esse protocolo buscou assegurar coerência entre os fundamentos teóricos, as escolhas técnicas e os procedimentos experimentais, funcionando como um guia para a seleção das abordagens de modelagem, das estratégias de treinamento e dos métodos de avaliação empregados ao longo da pesquisa.

Como etapa inicial, foi realizada uma pesquisa exploratória na literatura científica com o intuito de identificar abordagens, técnicas e estratégias computacionais utilizadas na previsão de doenças infecciosas. Essa etapa teve como finalidade mapear o estado da arte e compreender as principais linhas de investigação relacionadas à modelagem epidemiológica, ao uso de aprendizado de

máquina e à análise de séries temporais, servindo como subsídio para a definição do escopo metodológico da pesquisa.

A partir dessa análise preliminar, foram definidos termos-chave representativos do domínio investigado, contemplando expressões em língua portuguesa e inglesa. Com base nesses termos, foram elaboradas strings de busca genéricas, aplicadas em bases de dados científicas amplamente reconhecidas, tais como PubMed[1] e Capes Periódicos[2]. O Google Scholar[3] foi utilizado como ponto de partida, em função de sua ampla indexação, contribuindo tanto para o refinamento das palavras-chave quanto para a identificação das bases mais relevantes para a condução das buscas. As bases consultadas e as respectivas strings de busca utilizadas nessa etapa são apresentadas no Quadro 1.

Quadro 1 – Bases de dados e expressões gerais de busca

| Base de dados | String de busca |
|---|---|
| PubMed | *((machine learning[Title/Abstract] OR artificial intelligence[Title/Abstract])*<br>*AND (prediction[Title/Abstract] OR forecasting[Title/Abstract])*<br>*AND ((infectious disease[Title/Abstract] OR infectious diseases[Title/Abstract])*<br>*AND (outbreak[Title/Abstract] OR epidemic[Title/Abstract]))* |
| Capes Periódicos | *(“machine learning” OR “artificial intelligence” OR “deep learning”)*<br>*AND (“infectious disease” OR “infectious diseases” OR epidemic OR outbreak)*<br>*AND (prediction OR forecasting OR “early warning”)* |

**Fonte:** Autores**.**

As buscas realizadas retornaram um conjunto mais amplo de publicações relacionadas à modelagem preditiva de doenças infecciosas. Contudo, foram aplicados critérios de inclusão e exclusão com o objetivo de selecionar estudos metodologicamente comparáveis ao problema investigado neste trabalho.

Como critérios de inclusão e exclusão, consideraram-se:

[1] https://pubmed.ncbi.nlm.nih.gov/
[2] https://www.periodicos.capes.gov.br/index.php/acervo/lista-a-z-periodicos.html
[3] https://scholar.google.com.br/

Tabela 1 – Critérios de Inclusão

| ID | Critério de Inclusão |
|---|---|
| CI001 | Estudos que aplicam técnicas de aprendizado supervisionado ou modelos estatísticos à previsão de doenças infecciosas |
| CI002 | Trabalhos com descrição clara dos algoritmos utilizados |
| CI003 | Publicados entre 2020 e 2025. |
| CI004 | Pesquisas com foco explícito em previsão de casos ou detecção de surtos |
| CI005 | Artigos de acesso aberto. |
| CI006 | População-alvo composta por humanos. |

**Fonte:** Autores**.**

Tabela 2 – Critérios de Exclusão

| ID | Critério de Exclusão |
|---|---|
| CE001 | Estudos puramente teóricos, sem aplicação empírica ou validação experimental |
| CE002 | Trabalhos baseados exclusivamente em modelos mecanicistas sem componente orientado a dados |
| CE003 | Artigos sem detalhamento metodológico suficiente |
| CE004 | Pesquisas cujo contexto não esteja relacionado à vigilância epidemiológica |

**Fonte:** Autores**.**

Após a aplicação desses filtros, foram selecionados quatro estudos representativos, apresentados no Quadro 2, contemplando tanto abordagens baseadas em regressão quanto em classificação, permitindo uma análise comparativa consistente entre diferentes estratégias de modelagem epidemiológica.

Quadro 2 – Trabalhos relacionados em previsão e detecção de surtos epidemiológicos

| Autor | Contexto | Objetivo | Abordagem | Modelos | Métricas |
|---|---|---|---|---|---|
| Du & Pang (2021) | Influenza | Detectar e prever a ocorrência de surtos. | Classificação | Support Vector Machine, Gaussian Naive Bayes | Recall, Precision, Accuracy |
| Gao *et al*. (2024) | COVID-19 e SARS | Prever a ocorrência (ou ausência) de surtos antes que eles ocorram. | Classificação | Gradient Boosting, Logistic Regression, K-Nearest Neighbors, Support Vector Machine | AUC e Accuracy |
| Zanardo *et al*. (2024) | Dengue | Prever o número de casos semanais de dengue. | Regressão | Support Vector Machines, Random Forest, Gradient Boosting | MAE, RMSE, MASE, RMSSE, BIAS |
| Cabrera *et al*. (2022). | Dengue | Investigar a eficiência de modelos para prever o número de casos semanais de dengue. | Regressão e Classificação | Linear Regression, LASSO, Ridge, Support Vector Machine, Random Forest, Gradient Boosting, XGBoost, LightGBM | RMSE, MAE, $R^2$, Accuracy, Sensitivity, Specificity |

**Fonte:** Autores**.**

Com base na análise desses trabalhos, foi possível identificar algumas tendências recorrentes na literatura, como a predominância de modelos baseados em ensemble (especialmente *Random Forest* e *Gradient Boosting*), o uso frequente de métricas agregadas como *Accuracy* e AUC para avaliação de desempenho e a formulação crescente do problema como tarefa de classificação de risco, em vez de regressão estritamente numérica.

Ao mesmo tempo, foram observadas limitações importantes, como a sensibilidade de modelos de regressão a variáveis instáveis e flutuações abruptas, a

baixa ênfase na análise de incerteza das previsões e, em alguns casos, a ausência de estratégias adequadas de validação temporal.

Essas constatações orientaram a definição do escopo experimental da pesquisa, direcionando o foco para a implementação comparativa de modelos representativos, a adoção de validação temporal consistente e a incorporação da incerteza como critério adicional de avaliação em dados epidemiológicos reais, respeitando suas características temporais e regionais.

Posto isso, o estudo concentrou-se na etapa experimental, contemplando a preparação e organização dos dados, a implementação dos modelos computacionais, a definição das estratégias de treinamento e validação e a seleção das métricas de avaliação. O processo metodológico adotado buscou assegurar a comparabilidade entre as abordagens analisadas, a reprodutibilidade dos experimentos e a análise crítica dos resultados obtidos.

Dessa forma, o protocolo de pesquisa estabelecido conectou o levantamento conceitual inicial às etapas práticas de modelagem e avaliação, funcionando como um guia para a condução sistemática do estudo. Essa abordagem possibilitou integrar o conhecimento consolidado na literatura científica à aplicação experimental das técnicas investigadas, garantindo alinhamento metodológico com os objetivos propostos e sustentando a análise comparativa dos modelos preditivos desenvolvidos.

## 3.2 REGRESSÃO E O IMPACTO DE VARIÁVEIS INSTÁVEIS

Para Zanardo *et al.* (2024), que investigam a previsão semanal de casos de dengue em unidades federativas brasileiras a partir de séries temporais históricas e múltiplos preditores, o objetivo central de seu trabalho é estimar quantitativamente o número de casos futuros, utilizando modelos estatísticos e técnicas de aprendizado de máquina. No estudo, os autores ressaltam que a incidência da dengue é influenciada por uma combinação complexa de fatores, incluindo condições ambientais, características demográficas, circulação viral e possíveis interações com outras doenças infecciosas. Essa multiplicidade de fatores torna a dinâmica epidemiológica altamente não linear e sujeita a variações abruptas ao longo do tempo e do espaço.

Um ponto importante discutido por Zanardo *et al*. (2024) é a dificuldade de generalização dos modelos de regressão em diferentes contextos regionais. Mesmo

quando modelos multivariados são empregados, a relação entre variáveis explicativas e o número de casos não permanece estável ao longo dos anos, em especial durante períodos de surto. Fatores como mudanças climáticas atípicas, intervenções pontuais do poder público, variações na mobilidade populacional e alterações nos critérios de notificação impactam diretamente o processo gerador dos dados. Como consequência, modelos calibrados em determinados períodos tendem a apresentar erros significativos justamente nos momentos críticos, quando ocorre crescimento rápido da incidência.

Além disso, os autores avaliam o desempenho dos modelos utilizando métricas clássicas de regressão, como erro médio absoluto, raiz do erro quadrático médio e coeficiente de determinação. Embora essas métricas sejam adequadas para quantificar a precisão numérica das previsões, elas penalizam fortemente desvios em períodos de pico, que são característicos de surtos epidemiológicos. Assim, mesmo modelos com bom desempenho médio podem falhar em antecipar adequadamente situações de risco elevado, limitando sua utilidade operacional para vigilância epidemiológica.

De forma semelhante, o trabalho de Cabrera *et al*. compara métodos baseados em aprendizado de máquina para inferir o número de casos semanais de dengue também exemplifica a formulação do problema como regressão. O estudo analisa diferentes modelos e técnicas de pré-processamento, avaliando o desempenho por meio de métricas como MAPE, RMSE e $R^2$. Os resultados indicam que o desempenho dos modelos é altamente sensível à escolha das variáveis, à janela temporal considerada e às transformações aplicadas aos dados, evidenciando a instabilidade do problema.

Esse estudo reforça que, em séries epidemiológicas reais, a presença de ruído, atrasos de notificação e subnotificação compromete a capacidade dos modelos de regressão em fornecer previsões robustas e confiáveis. Em especial, a previsão do número exato de casos se mostra frágil em cenários de baixa incidência intercalados com picos abruptos, situação comum em doenças infecciosas. Dessa forma, ambos os trabalhos de regressão evidenciam que, embora a previsão quantitativa seja teoricamente desejável, ela enfrenta limitações estruturais impostas pela natureza instável e episódica dos surtos.

## 3.3 CLASSIFICAÇÃO E DETECÇÃO DE RISCO

Em contraste com a abordagem por regressão, Du e Pang propõem um método explicitamente formulado como um problema de classificação binária, cujo objetivo é prever se a próxima semana será caracterizada como surto ou não-surto. Os autores desenvolvem um indicador regional normalizado e utilizam esse indicador para rotular semanas históricas como pertencentes ou não a um surto, a partir de critérios estatísticos. A partir desses rótulos, o problema passa a ser tratado como uma tarefa de aprendizado supervisionado, em que a variável resposta assume valores discretos.

O modelo proposto combina classificadores como S*upport Vector Machine* e *Naive Bayes* Gaussiano em um esquema de ensemble, com o objetivo de aumentar a sensibilidade na detecção de surtos. Um aspecto relevante desse trabalho é o reconhecimento explícito do desbalanceamento dos dados, uma vez que semanas sem surto são muito mais frequentes do que semanas com surto. Por essa razão, os autores adotam o recall como métrica principal de avaliação, priorizando a capacidade do modelo de identificar corretamente surtos reais, mesmo ao custo de um aumento moderado de falsos positivos. As métricas *precision* e *accuracy* são utilizadas de forma complementar, sendo a acurácia tratada apenas como referência devido ao desbalanceamento das classes.

Essa escolha metodológica está fortemente alinhada aos objetivos da vigilância epidemiológica, na qual o custo de não detectar um surto pode ser significativamente maior do que o custo de emitir um alerta falso. O trabalho demonstra que a reformulação do problema como classificação permite maior robustez frente a ruídos e variações nos dados, além de produzir resultados mais diretamente acionáveis para gestores de saúde pública.

De forma complementar, Gao *et al*. apresentam um *framework* de classificação de séries temporais voltado à detecção precoce de surtos e não-surtos. Diferentemente de abordagens regressivas, os autores extraem um conjunto amplo de características estatísticas e indicadores de alerta precoce das séries de incidência, utilizando essas features como entrada para classificadores supervisionados. O objetivo não é prever o valor futuro da incidência, mas identificar padrões temporais que antecedem transições críticas no comportamento da série.

Os autores demonstram que essa abordagem é capaz de diferenciar estados epidemiológicos distintos com boa capacidade discriminativa, mesmo em cenários caracterizados por ruído e mudanças estruturais. Além disso, o trabalho reforça que a classificação de surtos permite maior flexibilidade na definição do horizonte de antecipação e na adaptação do modelo a diferentes contextos epidemiológicos. Assim como no trabalho de Du e Pang, as métricas de avaliação priorizam medidas de discriminação, como AUC, recall e precisão, em detrimento de métricas de erro numérico.

## 3.4 COMPARATIVO

No Quadro 2, a análise dos trabalhos sintetizados evidencia que a literatura recente em previsão e detecção de surtos epidemiológicos adota estratégias metodológicas distintas conforme o objetivo do estudo. Observa-se que abordagens baseadas em classificação são predominantemente utilizadas quando o interesse está na identificação antecipada de eventos críticos, como a ocorrência ou não de surtos, priorizando métricas associadas à capacidade discriminativa dos modelos, como *recall*, *precision* e AUC. Em contrapartida, os estudos que se propõem a estimar quantitativamente o número de casos recorrem a modelos de regressão e a métricas de erro contínuo, como MAE e RMSE, refletindo uma preocupação maior com a precisão numérica das previsões.

Entretanto, a comparação entre essas abordagens revela implicações relevantes para a aplicação prática dos modelos em contextos de vigilância em saúde. Trabalhos baseados em regressão tendem a apresentar maior sensibilidade a ruídos, atrasos de notificação e mudanças no processo gerador dos dados, especialmente em cenários epidemiológicos instáveis. Já as abordagens de classificação, ao operarem sobre estados epidemiológicos discretos, demonstram maior robustez frente a flutuações de curto prazo e maior alinhamento com processos decisórios operacionais. Além disso, observa-se que o aumento da complexidade dos modelos, com o uso de técnicas de ensemble e arquiteturas profundas, nem sempre se traduz em ganhos proporcionais de aplicabilidade, reforçando a importância de equilibrar desempenho preditivo, interpretabilidade e utilidade prática.

Desse modo, a análise dos trabalhos relacionados evidencia uma distinção conceitual clara entre prever quantitativamente o número de casos e classificar o risco de ocorrência de surtos. Os estudos baseados em regressão mostram que a previsão numérica é fortemente impactada por variáveis instáveis e por mudanças no processo gerador dos dados, o que compromete a confiabilidade das estimativas em períodos críticos. Em contrapartida, os trabalhos baseados em classificação demonstram que a identificação de estados epidemiológicos discretos é mais robusta, interpretável e alinhada às necessidades operacionais da vigilância em saúde.

Nesse contexto, a abordagem adotada neste trabalho, que classifica o risco de surto em níveis como normal, atenção e surto, entende-se como uma extensão natural das propostas encontradas na literatura recente. Ao introduzir uma classe intermediária de atenção, busca-se fornecer maior granularidade na avaliação do risco, permitindo ações graduais e proporcionais à gravidade esperada da situação epidemiológica. Essa reformulação preserva a capacidade de antecipação, ao mesmo tempo em que reduz a sensibilidade do modelo a flutuações numéricas de curto prazo.

# 4 SOLUÇÃO PROPOSTA

Neste capítulo é apresentada a solução proposta para apoiar a previsão da evolução temporal de doenças infecciosas por meio de técnicas de modelagem computacional e aprendizado de máquina. A solução foi concebida com o objetivo de explorar e comparar diferentes abordagens preditivas, aplicadas a dados epidemiológicos e demográficos, de modo a gerar estimativas capazes de auxiliar a análise epidemiológica e o apoio à tomada de decisão em saúde pública.

A proposta fundamenta-se na necessidade de dispor de métodos computacionais que possibilitem a análise sistemática de grandes volumes de dados em contextos epidemiológicos dinâmicos. A heterogeneidade das bases de dados disponíveis, aliada à complexidade dos fatores que influenciam a transmissão de doenças infecciosas, exige abordagens capazes de lidar com variáveis temporais e populacionais de forma integrada. Nesse contexto, a solução proposta busca estabelecer um processo organizado e reprodutível para o desenvolvimento, treinamento e avaliação de modelos preditivos, permitindo a análise comparativa de diferentes técnicas.

Desse modo, a solução foi concebida para atender pesquisadores e profissionais da área da saúde pública interessados em compreender e acompanhar a dinâmica de transmissão de doenças infecciosas em nível regional. Assim, o foco não se restringe à obtenção de previsões, mas inclui a transparência metodológica, a análise crítica do desempenho dos modelos e a interpretação dos resultados, aspectos essenciais para a utilização das estimativas em processos decisórios baseados em evidências.

Do ponto de vista técnico, a solução estrutura-se como um fluxo de etapas que compreende a preparação e organização dos dados, a implementação de diferentes modelos computacionais, a definição de estratégias de treinamento e validação e a avaliação do desempenho preditivo. Essa organização permite que cada etapa do processo seja analisada e ajustada de forma independente, favorecendo a adaptação da abordagem a diferentes conjuntos de dados e cenários epidemiológicos.

Por fim, neste capítulo são descritos os principais aspectos da solução, iniciando pela apresentação do pipeline de dados e das estratégias de modelagem adotadas. Em seguida, são detalhados os procedimentos utilizados para a avaliação comparativa dos modelos preditivos e a discussão de sua aplicabilidade no contexto da vigilância em saúde pública.

## 4.1 *PIPELINE* DE DADOS

Em contextos que envolvem grandes volumes de dados heterogêneos, como os dados epidemiológicos de sistemas nacionais de vigilância, o uso de um *pipeline* estruturado é fundamental para garantir rastreabilidade, reprodutibilidade e qualidade das informações utilizadas nas análises.

Assim, o *pipeline* de dados foi concebido para lidar com bases públicas de saúde caracterizadas por diferentes formatos, granularidades e níveis de qualidade. Portanto, sua função central é organizar o fluxo de dados desde a ingestão dos arquivos brutos provenientes do SINAN até a disponibilização de conjuntos de dados limpos e estruturados, adequados para a modelagem computacional e a aplicação de técnicas de inteligência artificial [OBJ].

O pipeline adotado está diretamente associado a uma arquitetura em camadas do tipo *Data Lake*, permitindo a separação clara entre dados brutos, dados transformados e dados prontos para análise. Essa organização facilita tanto o controle das transformações aplicadas quanto a reutilização dos dados em diferentes etapas do processo analítico. Desse modo, configurou-se a pipeline conforme descrito na Figura 3.

Figura 3 – Arquitetura do pipeline de dados adotado no estudo


**Fonte:** Autores.

A implementação da arquitetura de Data Lake foi realizada em nível conceitual e lógico, estruturada por meio da organização sistemática de diretórios e arquivos no ambiente local de desenvolvimento, distribuídos nas camadas Raw, Bronze, Silver e Gold. Cada camada foi organizada hierarquicamente por agravo, ano e unidade federativa, garantindo rastreabilidade e controle das transformações aplicadas ao longo do ciclo de vida dos dados.

A ingestão dos dados do SINAN foi realizada via API do DATASUS, com apoio da biblioteca PySUS para leitura de arquivos no formato DBC. A conversão para formatos tabulares manipuláveis (DBF e CSV) foi conduzida em Python, preservando os dados originais na camada Raw.

O processamento e transformação dos dados foram realizados com as bibliotecas Pandas e NumPy. Para enriquecimento demográfico, foram consumidas APIs públicas do IBGE e do SIDRA, por meio da biblioteca Requests, permitindo a incorporação de informações de população e área territorial. Também foram utilizados dados geoespaciais no formato shapefile, manipulados com suporte da biblioteca GeoPandas, para integração territorial.

A arquitetura respeita os princípios fundamentais do paradigma Data Lake, incluindo segregação de camadas, preservação dos dados brutos, rastreabilidade das transformações e preparação estruturada para consumo analítico na etapa de modelagem. O desenvolvimento foi conduzido em ambiente Jupyter Notebook, com scripts auxiliares em Python para automação de tarefas de atualização e integração de dados, garantindo reprodutibilidade e rastreabilidade do fluxo analítico.

### 4.1.1 Ingestão de dados

A etapa de ingestão de dados corresponde ao primeiro estágio do pipeline e tem como objetivo coletar os dados epidemiológicos em seu formato original, preservando ao máximo suas características iniciais. No projeto, os dados foram obtidos a partir do Sistema de Informação de Agravos de Notificação (SINAN), disponibilizados pelo DATASUS em formato DBC[4],.

[4] Trata-se de um formato compactado específico do sistema de dados do DATASUS, descrito em http://w3.datasus.gov.br/sia/index.php?area=01

A ingestão foi realizada de forma automatizada por meio de requisição à API oficial do DATASUS, permitindo o download sistemático dos arquivos referentes a diferentes doenças, anos e unidades federativas. Nesse estágio, não foram aplicados filtros ou tratamentos nos dados, uma vez que o objetivo principal foi garantir a coleta integral dos registros disponíveis, possibilitando uma compreensão inicial de sua estrutura e conteúdo.

Os arquivos obtidos foram organizados de acordo com o tipo de agravo e o ano de notificação, mantendo um padrão de nomenclatura que facilita a rastreabilidade e o gerenciamento dos dados ao longo das etapas subsequentes do pipeline.

### 4.1.2 Arquitetura de *Data Lake*

Após a ingestão, os dados passaram a ser organizados segundo uma arquitetura em camadas baseada no conceito de *Data Lake.* Essa abordagem foi adotada para estruturar o *pipeline* de dados de forma escalável e modular, permitindo o armazenamento de dados em diferentes níveis de processamento.

A arquitetura implementada é composta pelas camadas *Raw, Bronze, Silver* e *Gold*. A camada *Raw* armazena os dados brutos exatamente como foram obtidos do SINAN, preservando os arquivos no formato original. Na camada *Bronze*, os dados passam por processos iniciais de conversão, como a transformação dos arquivos DBC para DBF e, posteriormente, para CSV, viabilizando sua manipulação em estruturas tabulares.

A camada *Silver* concentra os dados já consolidados e padronizados, incluindo a unificação dos registros por estado e a incorporação de informações complementares, como o mapeamento de códigos de doenças e municípios. Por fim, a camada *Gold* reúne os dados analíticos finais, prontos para a aplicação das técnicas de modelagem preditiva e análise estatística.

Essa organização em camadas permite que cada estágio do pipeline seja analisado e validado de forma independente, além de facilitar a manutenção e a evolução da solução proposta.

### 4.1.3 Limpeza e Transformação dos dados

A etapa de limpeza dos dados teve como objetivo assegurar a consistência, confiabilidade e adequação do conjunto de dados para a modelagem computacional. No projeto, essa etapa foi aplicada principalmente sobre os dados organizados na camada *Silver* do *Data Lake*, após a consolidação dos registros referentes ao estado de Alagoas.

O processo de limpeza incluiu a remoção de registros duplicados com base em identificadores principais, garantindo que cada notificação representasse um evento único. Além disso, registros com valores ausentes em campos considerados críticos foram descartados, reduzindo a presença de inconsistências que poderiam comprometer o desempenho dos modelos preditivos.

Também foram realizadas transformações estruturais nos dados, incluindo a remoção de colunas irrelevantes para a análise, tais como variáveis clínicas específicas de cada agravo (campos de hospitalização, resultados laboratoriais, dados de bloqueio epidemiológico) e campos redundantes ou não utilizados no escopo do estudo.

Além disso, ainda no âmbito de transformações, a padronização envolveu a normalização do código do município para seis dígitos, uma vez os dados históricos contemplam a década passada e a nova padronização é recente, a conversão de datas para formato *datetime* e a obtenção dos nomes oficiais dos municípios via integração com a API do IBGE.

Quanto às variáveis derivadas, destaca-se a criação da coluna DOENCA_NOME, que mapeia os códigos das doenças para seus nomes descritivos, e a definição da variável CASO_CONFIRMADO, obtida a partir da recodificação do campo CLASSI_FIN em uma variável binária que indica a confirmação diagnóstica do caso.

Esses procedimentos resultaram em um conjunto de dados mais homogêneo e adequado para as etapas subsequentes do pipeline, contribuindo para que as análises e previsões realizadas se aproximassem do comportamento epidemiológico observado nos dados históricos.

## 4.2 DESENVOLVIMENTO DA SOLUÇÃO

O desenvolvimento da solução proposta foi conduzido a partir da aplicação prática dos princípios metodológicos definidos anteriormente, contemplando a preparação dos dados, a definição das estratégias de modelagem e a avaliação comparativa dos modelos preditivos. O foco principal dessa etapa consistiu na construção de um processo reprodutível de classificação do risco de surtos epidemiológicos, considerando a natureza temporal dos dados, o desbalanceamento entre as classes e a necessidade de priorizar a identificação de cenários críticos no contexto da vigilância em saúde.

A implementação da solução foi realizada na linguagem Python[5], utilizando bibliotecas amplamente consolidadas no ecossistema de ciência de dados. Para manipulação e tratamento dos dados foram empregadas as bibliotecas Pandas[6] e NumPy[7]. A modelagem e o treinamento dos algoritmos de classificação foram conduzidos com o auxílio da biblioteca Scikit-learn[8], incluindo os módulos de validação cruzada temporal (*TimeSeriesSplit*[9]), busca em grade (*GridSearchCV*[10]) e técnicas de balanceamento de classes, como *SMOTE*. Para a detecção de anomalias foram utilizados os algoritmos *Isolation Forest* e *Local Outlier Factor*, também disponíveis no *Scikit-learn*. A geração de gráficos e visualizações foi realizada com *Matplotlib*[11] e *Seaborn*[12]. O desenvolvimento foi conduzido em ambiente *Jupyter Notebook*[13], permitindo rastreabilidade das etapas e reprodutibilidade dos experimentos.

A solução foi desenvolvida com base em dados epidemiológicos abrangendo o período de 2014 a 2024. A definição desse intervalo temporal fundamenta-se em critérios epidemiológicos relacionados à análise de tendências e à modelagem de séries temporais em saúde pública. Conforme discutido por Antunes e Cardoso (2015), a utilização de séries históricas extensas é essencial

---

[5] https://www.python.org/
[6] https://pandas.pydata.org/
[7] https://numpy.org/
[8] https://scikit-learn.org/
[9] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html
[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[11] https://matplotlib.org/
[12] https://seaborn.pydata.org/
[13] https://jupyter.org/

para a identificação consistente de padrões sazonais, tendências de longo prazo e variações interanuais nos dados epidemiológicos. Nesse contexto, a adoção de uma década de observação possibilita capturar múltiplas repetições dos ciclos sazonais e reduzir o impacto de oscilações pontuais ou anos atípicos, contribuindo para maior estabilidade estatística das estimativas. Séries temporais curtas podem não representar adequadamente o comportamento histórico do agravo, dificultando a distinção entre mudanças estruturais e flutuações ocasionais.

Conforme as Figuras 4 e 5, é possível observar como os dados foram recebidos, explicitados nos pontos abaixo:

- As colunas não têm nomes explícitos, sendo necessário conhecimento prévio ou consulta à documentação do SINAN[14];
- Há codificação em algumas colunas, como Município, Agravo, Doença, Classificação Final;
- Os dados estão organizados por semana e ano;
- Nem todas as colunas possuem dados;
- Nem todos os casos foram encerrados.

Figura 4 – Exemplo 1 de dados brutos

| | TP_NOT | ID_AGRAVO | DT_NOTIFIC | SEM_NOT | NU_ANO | SG_UF_NOT | ID_MUNICIP | ID_REGIONA | NU_IDADE_N |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 2 | G039 | 2014-01-01 | 201401 | 2014 | 27 | 270430 | 1533.0 | 3011 |
| 1 | 2 | B019 | 2014-01-01 | 201401 | 2014 | 27 | 270940 | 1542.0 | 4005 |
| 2 | 2 | A379 | 2014-01-02 | 201401 | 2014 | 27 | 270430 | 1533.0 | 4008 |
| 3 | 2 | A379 | 2014-01-02 | 201401 | 2014 | 27 | 270430 | 1533.0 | 4008 |
| 4 | 2 | A379 | 2014-01-02 | 201401 | 2014 | 27 | 270430 | 1533.0 | 3002 |
| 5 | 2 | B019 | 2014-01-02 | 201401 | 2014 | 27 | 270430 | 1533.0 | 4005 |
| 6 | 2 | B19 | 2014-01-03 | 201401 | 2014 | 27 | 270430 | 1533.0 | 4017 |
| 7 | 2 | B19 | 2014-01-03 | 201401 | 2014 | 27 | 270250 | 1540.0 | 4011 |
| 8 | 2 | A379 | 2014-01-03 | 201401 | 2014 | 27 | 270430 | 1533.0 | 4006 |
| 9 | 2 | A379 | 2014-01-04 | 201401 | 2014 | 27 | 270430 | 1533.0 | 2026 |

**Fonte:** Autores**.**

14 https://portalsinan.saude.gov.br/images/documentos/Agravos/Notificacao_Individual/DIC_DADOS_NET---Notificao-Individual_rev.pdf

Figura 5 – Exemplo 2 de dados brutos

| CS_SEXO | CS_GESTANT | CS_ESCOL_N | SG_UF | ID_MN_RESI | ID_RG_RESI | ID_PAIS | CLASSI_FIN | DT_ENCERRA | ID_DOENCA |
|---|---|---|---|---|---|---|---|---|---|
| M | 6.0 | 10.0 | 27 | 270690 | 1533 | 1 | 1.0 | 2014-01-09 | MENIBR |
| M | 6.0 | 10.0 | 27 | 270430 | 1533 | 1 | NaN | NaN | VARCBR |
| F | 6.0 | 2.0 | 27 | 270430 | 1533 | 1 | 1.0 | 2014-02-25 | COQUBR |
| M | 6.0 | NaN | 27 | 270430 | 1533 | 1 | 1.0 | 2014-02-25 | COQUBR |
| M | 6.0 | 10.0 | 27 | 270080 | 1541 | 1 | 1.0 | 2014-03-02 | COQUBR |
| M | 6.0 | 10.0 | 27 | 270430 | 1533 | 1 | 1.0 | 2014-01-12 | VARCBR |
| M | 6.0 | 4.0 | 27 | 270430 | 1533 | 1 | 1.0 | 2014-02-17 | HEPABR |
| M | 6.0 | 1.0 | 27 | 270250 | 1540 | 1 | 1.0 | 2014-04-30 | HEPABR |
| M | 6.0 | 10.0 | 27 | 270430 | 1533 | 1 | 1.0 | 2014-02-25 | COQUBR |
| F | 6.0 | 10.0 | 27 | 270770 | 1533 | 1 | 1.0 | 2014-02-13 | COQUBR |

**Fonte:** Autores**.**

Além disso, conforme descrito anteriormente, o intervalo correspondente à pandemia de COVID-19 causou um impacto significativo na notificação dos casos e, por isso, foi excluído da análise, conforme será detalhado posteriormente.

Sendo assim, a partir do conjunto de dados obtidos, foram implementadas estratégias de definição de variáveis e classes, seleção de hiperparâmetros, tratamento de valores atípicos e treinamento dos modelos de classificação, descritas nas subseções a seguir.

### 4.2.1 Definição de Variáveis e Classes

A definição das variáveis de entrada e da variável-alvo constituiu uma etapa central no desenvolvimento da solução, uma vez que determina diretamente a capacidade dos modelos em representar o fenômeno epidemiológico analisado. As variáveis de entrada foram construídas a partir dos dados epidemiológicos disponíveis, incorporando informações temporais e populacionais relevantes para a caracterização da dinâmica de transmissão das doenças infecciosas.

Entre as *features* utilizadas destacam-se o número absoluto de casos notificados por período, defasagens temporais (*lags*), médias móveis, variações percentuais entre períodos consecutivos e taxas de incidência ajustadas pela população. Essas transformações foram adotadas com o objetivo de capturar padrões de tendência, sazonalidade e mudanças abruptas na evolução temporal dos casos, conforme discutido na literatura epidemiológica e em estudos recentes baseados em séries temporais.

A definição das classes de risco epidemiológico fundamenta-se no conceito de linha de base histórica e de canal endêmico, amplamente utilizado na vigilância epidemiológica para identificação de desvios do comportamento esperado de um agravo ao longo do tempo. Conforme descrito no Guia de Vigilância em Saúde do Ministério da Saúde (BRASIL, 2022), a detecção de surtos envolve a comparação entre o número observado de casos e parâmetros históricos previamente estabelecidos, permitindo a identificação de aumentos acima do padrão esperado.

Nesse contexto, foi adotado um critério estatístico baseado na média histórica acumulada e no desvio padrão da série temporal até o período imediatamente anterior ao mês analisado. Essa abordagem está alinhada aos princípios da análise de séries temporais aplicadas à epidemiologia, nos quais a média histórica representa o comportamento esperado do agravo, enquanto o desvio padrão expressa a variabilidade natural do fenômeno (ANTUNES; CARDOSO, 2015).

A variável-alvo foi definida como uma classificação discreta do risco epidemiológico, a partir de um critério estatístico aplicado à série histórica de casos mensais. Para cada mês analisado, foram calculadas a média e o desvio padrão históricos de forma expansiva até o mês imediatamente anterior, evitando o uso de informações futuras no processo de rotulagem.

Seja $Y_t$ o número de casos observados no período *t*, define-se a média histórica expansiva até o período anterior, como mostrado na Equação (1):

$$\mu_t = \frac{1}{t-1} \sum_{i=1}^{t-1} Y_i \tag{2}$$

E o desvio padrão histórico conforme a Equação (2):

$$\sigma_t = \sqrt{\frac{1}{t-2} \sum_{i-1}^{t-1} (Y_i - \mu_t)^2} \tag{3}$$

Com base nesses limiares, foram definidas três classes de risco:

- Normal: quando o número de casos do mês é inferior ou igual à média histórica, ou seja, se $Y_t \leq \mu_t$;
- Atenção: quando o número de casos situa-se entre a média histórica e a média acrescida de um desvio padrão, ou seja, se $\mu_t < Y_t \leq \mu_t + \sigma_t$; e
- Surto: quando o número de casos excede a média histórica acrescida de um desvio padrão, ou seja, se $Y_t > \mu_t + \sigma_t$.

Figura 6 – Exemplo de aplicação do critério estatístico para classificação do risco epidemiológico

| Doença | Casos | Média Histórica | Desvio Padrão | Classe |
|---|---|---|---|---|
| Coqueluche | 27 | 21.33 | 13.38 | Atenção |
| Hepatite | 12 | 11.53 | 5.33 | Atenção |
| Meningite | 14 | 11.2 | 3.12 | Atenção |
| Doenças Exantemáticas | 12 | 0.67 | 0.9 | Surto |
| Varicela (Catapora) | 9 | 8.5 | 2.9 | Atenção |
| Doenças Exantemáticas | 0 | 1.5 | 3.19 | Normal |
| Coqueluche | 0 | 7.47 | 9.87 | Normal |

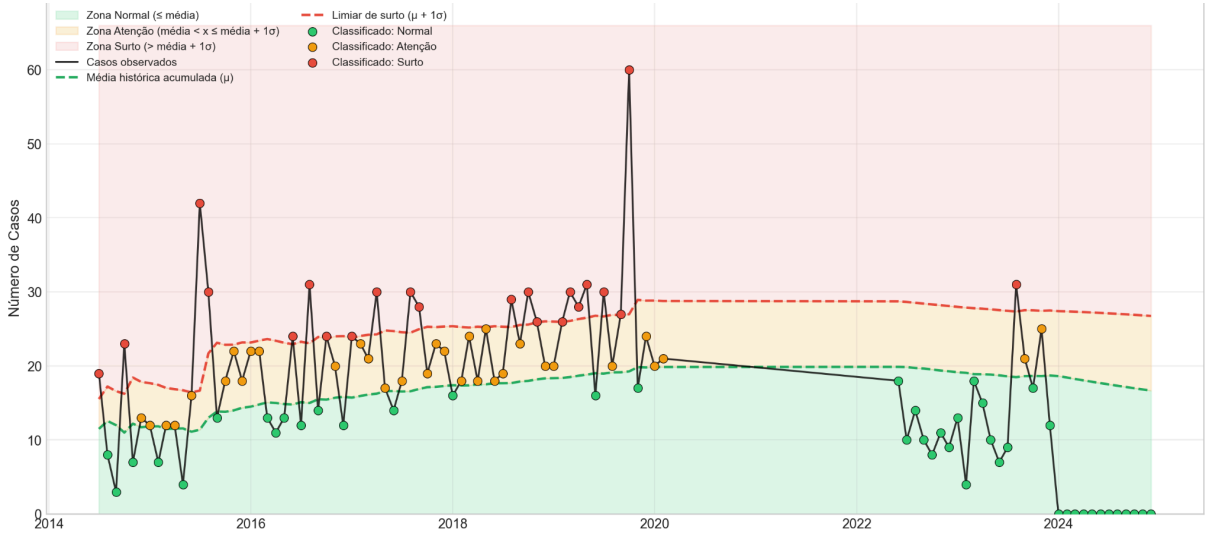
**Fonte:** Autores**.**

Todavia, essa definição tende a ser confusa, uma vez que o próprio *dataset* do SINAN traz registros com Tipo de Notificação (tp_notificacao), que varia entre os valores: (1) Negativa, (2) Individual, (3) Surto e (4) Agregado. Porém, esses valores não se refletem nos dados, pois ao analisar todo o conjunto de informações obtidas, não foram encontradas nenhum valor além do (2) Individual, ainda que seja possível notar picos de surtos ao expor graficamente os dados.

Desse modo, essa formulação da classe Risco permite identificar estados epidemiológicos distintos de maneira objetiva e reprodutível, ao mesmo tempo em que introduz uma classe intermediária capaz de capturar elevações relevantes sem caracterizar imediatamente um surto. Ao formular o problema como uma tarefa de classificação, busca-se reduzir a sensibilidade do modelo a flutuações numéricas de curto prazo e alinhar os resultados às necessidades operacionais da vigilância em saúde pública.

Na figura 7, é possível notar toda a variação da classificação de risco ao longo dos 10 anos, para Hepatite, em Maceió, que demonstra picos epidemiológicos bem definidos no início do ano de 2015 e no final do ano de 2019.

Figura 7 – Regra de Classificação de Risco para Hepatite em Maceió




**Fonte:** Autores.

### 4.2.2 Seleção de Hiperparâmetros

A seleção de hiperparâmetros dos modelos de classificação foi realizada por meio de busca em grade (*grid search*), associada a um esquema de validação cruzada temporal. Essa escolha decorre do fato de que os dados utilizados apresentam dependência temporal, tornando inadequadas abordagens de validação baseadas em amostragem aleatória. Para esse fim, foi adotado o método *TimeSeriesSplit*, que preserva a ordem cronológica dos dados ao longo dos diferentes conjuntos de treinamento e validação.

Como métrica de otimização durante a busca de hiperparâmetros, foi utilizado o *F1-score macro*, que considera de forma equilibrada o desempenho do modelo em todas as classes de risco (Normal, Atenção e Surto). Essa escolha evita que o processo de otimização privilegie exclusivamente a classe majoritária, garantindo maior sensibilidade à classe Surto, que possui maior relevância operacional no contexto da vigilância epidemiológica.

Foram avaliadas diferentes famílias de modelos de aprendizado de máquina, incluindo Regressão Logística, *Random Forest* e *Gradient Boosting*. Para

cada abordagem, foram definidos conjuntos de hiperparâmetros compatíveis com suas características, como profundidade máxima das árvores, número de estimadores, taxa de aprendizado e pesos de classe. As tabelas 3, 4 e 5 demonstram o que foi utilizado para cada modelo.

Tabela 3 – Hiperparâmetros de Random Forest

| Hiperparâmetro | Valor |
|---|---|
| n_estimators | 200 |
| max_depth | 10 |
| min_samples_split | 2 |
| min_samples_leaf | 1 |
| class_weight | balanced_subsample |
| random_state | 42 |

**Fonte:** Autores.

Tabela 4 – Hiperparâmetros de Logistic Regression

| Hiperparâmetro | Valor |
|---|---|
| C | 1 |
| penalty | l2 |
| solver | lbfgs |
| class_weight | balanced |
| max_iter | 2000 |
| random_state | 42 |

**Fonte:** Autores.

Tabela 5 – Hiperparâmetros de Gradient Boosting

| Hiperparâmetro | Valor |
|---|---|
| n_estimators | 200 |
| max_depth | 7 |
| learning_rate | 0.05 |
| min_samples_split | 5 |
| min_samples_leaf | 2 |
| loss | log_loss |
| criterion | friedman_mse |
| subsample | 1.0 |
| random_state | 42 |

**Fonte:** Autores.

A seleção do modelo final considerou não apenas o valor do *F1-score macro*, mas também a incerteza média das previsões, calculada a partir da entropia

das probabilidades preditas, buscando um compromisso entre desempenho preditivo e confiabilidade das classificações.

### 4.2.3 *Outliers*

O tratamento de valores atípicos foi abordado de forma diferenciada, considerando tanto aspectos contextuais quanto padrões pontuais nos dados. Inicialmente, o período correspondente à pandemia de COVID-19, compreendido entre março de 2020 e maio de 2022, conforme definido pela Organização Mundial da Saúde, foi excluído da base de dados. Esse intervalo foi caracterizado como um outlier de contexto, uma vez que alterou de forma significativa os padrões de notificação e a dinâmica de transmissão de diversas doenças infecciosas, podendo comprometer a definição estatística de surto e o processo de aprendizado dos modelos.

Além dessa exclusão temporal, não foi realizada a remoção direta de observações pontuais consideradas atípicas. Em vez disso, adotou-se uma abordagem complementar de detecção de anomalias, com o objetivo de identificar padrões incomuns sem alterar o conjunto de dados utilizado no treinamento dos classificadores. Para essa análise, foram aplicados algoritmos como *Isolation Forest* e *Local Outlier Factor*, utilizando variáveis relacionadas ao número de casos, defasagens temporais, médias móveis, variações percentuais e taxas populacionais.

Essa análise permitiu comparar os resultados da detecção de anomalias com a classificação de risco de surto, evidenciando que nem todo surto estatístico corresponde a uma anomalia e que padrões atípicos podem ocorrer mesmo em períodos classificados como Normal ou Atenção. Dessa forma, a detecção de anomalias foi utilizada como uma camada adicional de interpretação, sem interferir diretamente no processo de treinamento dos modelos preditivos.

### 4.2.4 Treinamento

O treinamento dos modelos preditivos seguiu uma estratégia estruturada, composta pela divisão temporal dos dados, balanceamento das classes e otimização dos hiperparâmetros. Os dados foram ordenados cronologicamente e divididos em conjuntos de treinamento e teste, utilizando aproximadamente 80% das observações

mais antigas para treinamento e os 20% mais recentes para avaliação, simulando um cenário realista de previsão.

Especificamente para o Gradient Boosting, que trata-se de uma família de algoritmos com algumas variações entre suas implementações, o modelo utilizadofoi por meio da classe GradientBoostingClassifier.. Essa implementação utiliza árvores de decisão como estimadores base, construídas sequencialmente, de modo que cada nova árvore corrige os erros residuais das anteriores, justificando essa escolha devido à sua robustez, ampla validação na literatura e integração nativa com o ecossistema de validação cruzada e otimização de hiperparâmetros da própria biblioteca.

No contexto geral, devido ao desbalanceamento natural entre as classes, especialmente da classe Surto, foi aplicado o método *SMOTE* exclusivamente no conjunto de treinamento. Essa técnica permitiu aumentar a representatividade das classes minoritárias sem introduzir viés no conjunto de teste, que manteve a distribuição original dos dados, conforme a tabela 6.

Tabela 6 – Balanceamento de classes com SMOTE

| Classe | Antes | Depois |
|---|---|---|
| Normal | 122 | 122 |
| Atenção | 59 | 122 |
| Surto | 27 | 122 |
| Total | 208 | 366 |

**Fonte:** Autores**.**

Conforme a Tabela 6, foram geradas 158 amostras sintéticas, resultando em um aumento total de 76,0% no volume de dados de treinamento. O balanceamento garantiu que todas as classes apresentassem a mesma representatividade, reduzindo o viés do modelo em favor da classe majoritária. Com isso, essa etapa foi particularmente relevante, considerando que, no contexto da vigilância epidemiológica, erros associados à subdetecção da classe Surto podem gerar consequências operacionais mais graves do que falsos positivos.

Além das métricas tradicionais de desempenho, foi incorporada à solução a quantificação do grau de incerteza associado às previsões. Para cada instância classificada, o modelo gera probabilidades para as três classes de risco, a partir das

quais foi calculada a entropia normalizada da distribuição de probabilidades. Valores baixos indicam previsões mais confiáveis, enquanto valores elevados sinalizam maior incerteza na classificação.

O grau de incerteza foi utilizado tanto como critério auxiliar na seleção do modelo final quanto como elemento interpretativo das previsões. Modelos com bom desempenho preditivo, mas elevada incerteza média, foram preteridos em favor de abordagens mais estáveis. Nas previsões prospectivas, a incerteza associada a cada classificação permite indicar cenários que demandam maior cautela na interpretação, oferecendo subsídios adicionais para a tomada de decisão em vigilância epidemiológica.

# 5 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados obtidos a partir da aplicação da metodologia proposta, contemplando a avaliação comparativa dos modelos de classificação, a análise de desempenho e incerteza das previsões, a identificação de padrões anômalos e a geração de previsões prospectivas de risco epidemiológico para o município de Maceió - AL.

## 5.1 PRÉ PROCESSAMENTO DOS DADOS

Os dados utilizados neste estudo foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN), abrangendo o período de janeiro de 2014 a dezembro de 2024. A base original continha 7.865 registros individuais de notificações de doenças infecciosas no estado de Alagoas, distribuídos em 115 municípios. Esses registros representam eventos notificados no sistema de vigilância epidemiológica, refletindo tanto a incidência das doenças quanto a dinâmica de notificação e confirmação diagnóstica ao longo do tempo.

### 5.1.1 Estrutura dos Dados Brutos

Os registros individuais do SINAN eram compostos por 22 variáveis, organizadas em diferentes categorias de informação. Entre elas, destacam-se variáveis relacionadas à identificação da notificação (TP_NOT, ID_AGRAVO, DT_NOTIFIC), características demográficas do paciente (NU_IDADE_N, CS_SEXO, CS_GESTANT, CS_ESCOL_N), informações geográficas (ID_MUNICIP, SG_UF, ID_MN_RESI), classificação final do caso (CLASSI_FIN) e identificação do agravo (ID_DOENCA, DOENCA_NOME).

Essa estrutura evidencia que a base foi originalmente concebida para fins administrativos e de vigilância clínica, voltada ao registro individual de notificações e à consolidação estatística para monitoramento em saúde pública. Como consequência, os dados encontram-se organizados em formato transacional, com registros individuais e múltiplas variáveis categóricas, ausência de estrutura temporal agregada e presença de campos incompletos ou heterogêneos. Para sua utilização em modelos de aprendizado de máquina, especialmente em tarefas de

previsão temporal, foi necessária a transformação da base em séries agregadas, a padronização de variáveis e a definição de critérios explícitos de filtragem e confirmação diagnóstica, conforme descrito no capítulo 4..

A distribuição das doenças notificadas no período apresentou a seguinte composição:

- Hepatite (2.643 registros);
- Meningite (2.173);
- Varicela (1.776);
- Coqueluche (750);
- Doenças Exantemáticas (395);
- Paralisia Flácida Aguda (96);
- Tétano Acidental (16);
- Difteria (6);
- Rotavírus (2); e
- Raiva Humana (2).

Observa-se que a distribuição é significativamente desigual entre os agravos, com forte concentração em Hepatite e Meningite, enquanto outras doenças apresentam ocorrência residual. Essa assimetria tem implicações diretas na modelagem, especialmente no que se refere à estabilidade estatística das séries temporais e ao desbalanceamento entre classes.

### 5.1.2 Critérios de Filtragem e Seleção

Com o objetivo de maximizar a precisão dos modelos preditivos, o escopo geográfico foi delimitado ao município de Maceió, responsável por 63,0% das notificações estaduais (4.958 registros). Essa escolha fundamenta-se na necessidade de garantir volume amostral suficiente para construção de séries temporais consistentes, além de reduzir heterogeneidades regionais que poderiam introduzir ruídos adicionais na modelagem.

Assim, a filtragem dos dados ocorreu em três etapas principais: Confirmação diagnóstica, Relevância epidemiológica e Exclusão do período pandêmico.

Para a etapa de Confirmação diagnóstica, inicialmente, foram removidos registros sem data de notificação válida (DT_NOTIFIC), garantindo consistência temporal para a agregação posterior. Em seguida, aplicou-se o filtro geográfico para o município de Maceió - AL. A definição de caso confirmado não se restringiu a uma regra única para todas as doenças, sendo adotada uma lógica específica por agravo, conforme os critérios epidemiológicos de classificação final (CLASSI_FIN) do SINAN. Assim, nesse contexto, a Tabela 7 foi gerada.

Tabela 7 – Mapeamento da Classificação Final dos casos pelo SINAN

| Doença | Valores | Descrição dos códigos |
|---|---|---|
| Hepatite | 1, 2, 4 | 1 = Confirmação laboratorial;<br>2 = Confirmação clínico-epidemiológica;<br>4 = Cicatriz sorológica |
| Meningite | 1 | 1 = Confirmado |
| Coqueluche | 1 | 2 = Confirmado |
| Varicela (Catapora) | 1 | 3 = Confirmado |
| Doenças Exantemáticas | 1, 2 | 1 = Sarampo;<br>2 = Rubéola |

**Fonte:** Autores**.**

Em seguida, agora para Relevância epidemiológica, foram selecionadas apenas doenças com pelo menos 100 casos confirmados no período analisado. Esse limiar foi definido com o objetivo de evitar séries excessivamente esparsas, que comprometem tanto a definição estatística de surtos quanto o desempenho dos algoritmos de classificação. Permaneceram, assim, cinco agravos: Hepatite (1.985 casos), Meningite (1.802), Coqueluche (390), Varicela (346) e Doenças Exantemáticas (167).

Por fim, foi realizada a Exclusão do período pandêmico, compreendido entre março de 2020 e maio de 2022. A pandemia de COVID-19 provocou alterações significativas nos padrões de notificação e na dinâmica de transmissão das demais doenças infecciosas, seja por mudanças comportamentais da população, seja por reorganização dos serviços de saúde. A manutenção de dados referentes a esse intervalo poderia distorcer os limites estatísticos utilizados na classificação de risco e introduzir viés no treinamento dos modelos.

### 5.1.3 Agregação Temporal

Após a filtragem, os dados individuais foram agregados por doença e por mês, transformando registros clínicos individuais em séries temporais mensais. Essa agregação foi adotada para alinhar a estrutura dos dados ao objetivo do estudo, que consiste na previsão de risco epidemiológico em escala temporal mensal.

Para garantir consistência na estrutura temporal, foi construído um grid completo contendo todas as combinações possíveis de *doença* × *mês* dentro do período analisado, incluindo meses sem ocorrência de casos, preenchidos com valor zero. Essa etapa evita lacunas temporais e assegura que os modelos possam aprender tanto padrões de ocorrência quanto períodos de estabilidade.

Após a exclusão do intervalo pandêmico, a base agregada totalizou 525 observações mensais. Para cada observação foram computadas duas métricas principais: *o número de casos confirmados* e *o total de notificações registradas*, permitindo análises complementares entre incidência confirmada e volume bruto de registros.

### 5.1.4 Estatísticas Descritivas

A Tabela 8 apresenta as estatísticas descritivas dos casos mensais por doença no município de Maceió, considerando o período de 2014 a 2024, excluído o intervalo pandêmico.

Tabela 8 – Estatísticas descritivas dos casos notificados

| Doença | Total de Casos | Média Mensal | Desvio Padrão | Máximo Mensal |
|---|---|---|---|---|
| Hepatite | 1733 | 16,5 | 10,17 | 60 |
| Meningite | 1612 | 15,35 | 11,77 | 64 |
| Coqueluche | 380 | 3,62 | 6,25 | 47 |
| Varicela | 335 | 3,19 | 3,65 | 14 |
| Doenças Exantemáticas | 151 | 1,44 | 3,22 | 24 |

**Fonte:** Autores**.**

Na tabela acima, Hepatite e Meningite apresentaram as maiores médias mensais de casos, indicando maior persistência ao longo do tempo. Em contraste, Coqueluche demonstrou elevada variabilidade (coeficiente de variação de 172,7%), caracterizando um padrão episódico com picos concentrados em determinados

períodos e longos intervalos de baixa incidência. Varicela e Doenças Exantemáticas apresentaram médias menores e dispersão intermediária, sugerindo comportamento sazonal ou intermitente.

Essas diferenças estruturais entre as séries temporais reforçam a necessidade de abordagens de modelagem capazes de lidar com variabilidade heterogênea e eventos esporádicos de maior intensidade.

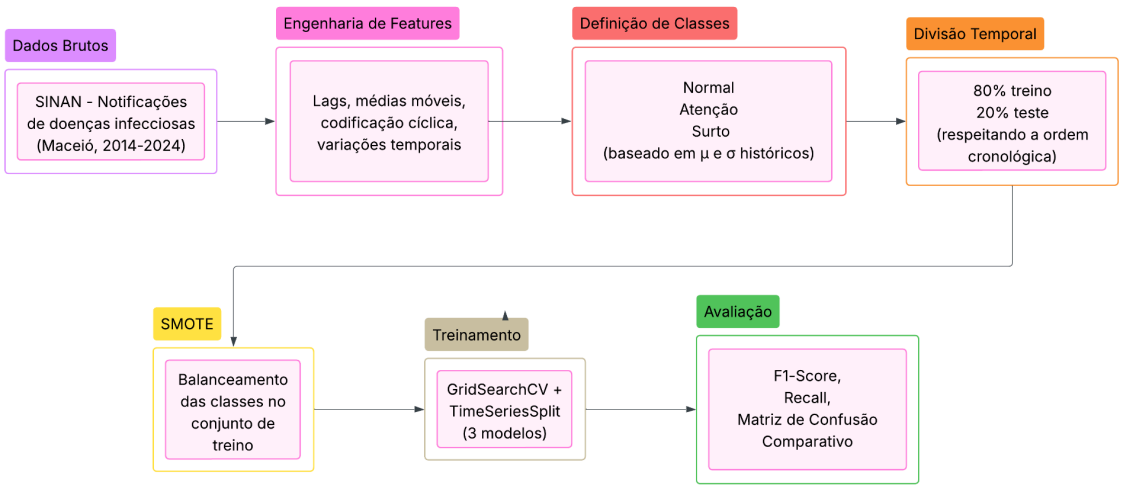
### 5.1.5 Estrutura Final dos Dados

Após o processamento e consolidação, a base agregada passou a conter as variáveis ID_MUNICIP, NOME_MUNICIP, DOENCA_NOME, ANO, MES, CASOS, NOTIFICACOES, DATA e SURTO, conforme descrita no capítulo 4.

A partir disso, foram derivadas variáveis temporais, defasagens (lags), médias móveis e limiares estatísticos empregados na classificação do risco de surto. Dessa forma, a organização final dos dados estabeleceu as condições necessárias para aplicação consistente das técnicas de aprendizado de máquina descritas nos capítulos seguintes.

### 5.1.6 Arquitetura final

Figura 8 – Arquitetura final da solução




**Fonte:** Autores.

A Figura 8 apresenta uma visão geral da arquitetura adotada para o treinamento dos modelos de classificação de risco de surto. O fluxograma sintetiza

as principais etapas do processo, desde a organização dos dados brutos provenientes do SINAN até a avaliação comparativa dos modelos. A estrutura foi concebida de modo a garantir coerência temporal, reprodutibilidade dos experimentos e adequação metodológica às características das séries epidemiológicas analisadas.

Logo, essa representação visual organiza as decisões metodológicas adotadas e evidencia a sequência lógica que orientou a implementação e a experimentação computacional descritas nas subseções seguintes.

## 5.2 AVALIAÇÃO COMPARATIVA DOS MODELOS DE CLASSIFICAÇÃO

A escolha dos modelos avaliados neste trabalho foi orientada por critérios metodológicos relacionados à interpretabilidade do modelo, robustez frente a ruído, capacidade de generalização e desempenho em tarefas de classificação epidemiológica, além de evidências da literatura, como descrito no Capítulo 3, que demonstram a aplicação recorrente dessas abordagens na previsão e detecção de surtos.

Conforme sintetizado no Quadro 2, estudos anteriores têm empregado predominantemente abordagens de classificação baseadas em modelos lineares e técnicas de *ensemble*, com destaque para algoritmos como Regressão Logística, *Random Forest* e *Gradient Boosting*. Dessa forma, optou-se por avaliar modelos representativos de diferentes níveis de complexidade e capacidade de generalização, permitindo uma comparação sistemática entre abordagens interpretáveis, robustas e de maior poder discriminativo.

A Regressão Logística foi incluída como modelo de referência, amplamente utilizada em estudos epidemiológicos por sua simplicidade e interpretabilidade. O *Random Forest* foi selecionado por sua robustez frente a ruído e por sua capacidade de capturar relações não lineares em conjuntos de dados heterogêneos. Já o *Gradient Boosting* foi escolhido por seu desempenho consistente em trabalhos recentes voltados à detecção de surtos, especialmente em cenários nos quais a identificação correta da classe crítica é prioritária. A avaliação comparativa desses modelos permite analisar não apenas o desempenho preditivo,

mas também os compromissos entre interpretabilidade, sensibilidade ao surto e estabilidade das previsões.

Tabela 9 – Modelos e Métricas

| Modelos | Métricas | | | | | |
|---|---|---|---|---|---|---|
| | Accuracy | F1 Macro | Recall (Surto) | Precision (Surto) | F1 (Surto) | Média de Incerteza |
| Random Forest | 0.8167 | 0.7569 | 0.7895 | 0.7143 | 0.7500 | 0.6590 |
| Gradient Boosting | 0.7833 | 0.7317 | 0.8421 | 0.7273 | 0.7805 | 0.0863 |
| Logistic Regression | 0.700 | 0.6287 | 0.6842 | 0.7222 | 0.7027 | 0.4152 |

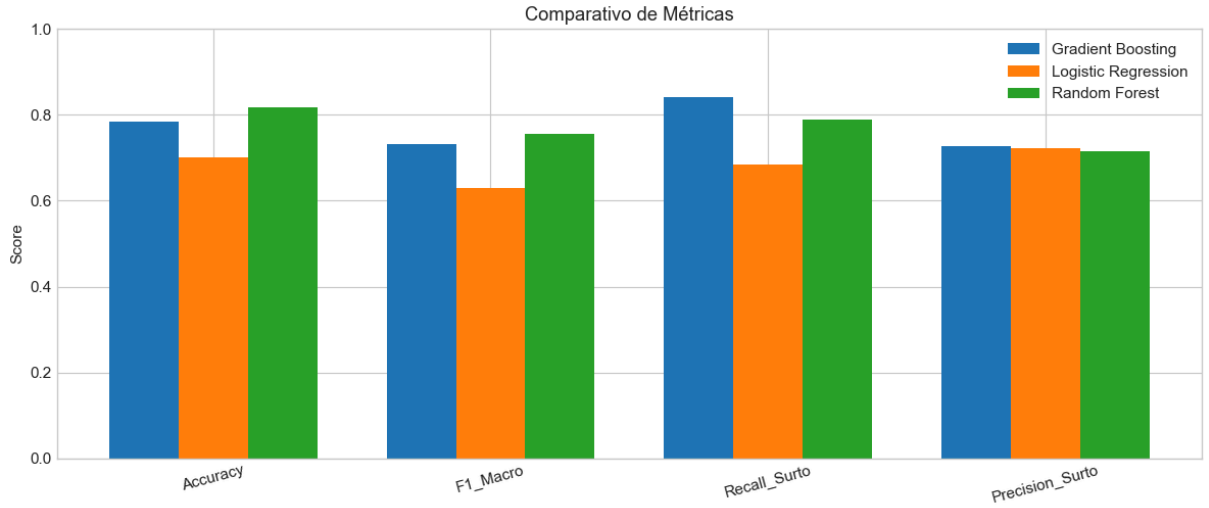
**Fonte:** Autores.
**Legenda:** Azul representa os melhores valores; Laranja representa os valores intermediários; e Vermelho representa os piores valores.

A Tabela 9 apresenta o desempenho dos modelos *Random Forest*, *Gradient Boosting* e Regressão Logística, considerando métricas globais e específicas para a classe Surto, além do grau médio de incerteza das previsões. Observa-se que o *Random Forest* apresentou a maior acurácia geral (0,8167) e o maior *F1-score macro* (0,7569), indicando bom equilíbrio no desempenho entre as classes. No entanto, esse modelo também apresentou a maior incerteza média (0,6590), sugerindo menor confiabilidade nas probabilidades preditas.

O *Gradient Boosting* destacou-se por apresentar o maior recall para a classe Surto (0,8421), métrica fundamental no contexto da vigilância epidemiológica, pois indica maior capacidade de identificar corretamente cenários críticos. Além disso, esse modelo apresentou a menor incerteza média entre os avaliados (0,0863), evidenciando maior estabilidade e confiança nas previsões. Além disso, obteve valores muito próximos do *Random Forest* para acurácia e F1-score, o que indica bom desempenho.

A Regressão Logística, por sua vez, apresentou desempenho inferior nas métricas globais e maior incerteza relativa, indicando limitações para capturar padrões mais complexos presentes nos dados epidemiológicos.

Figura 9 – Gráfico Comparativo de Métricas



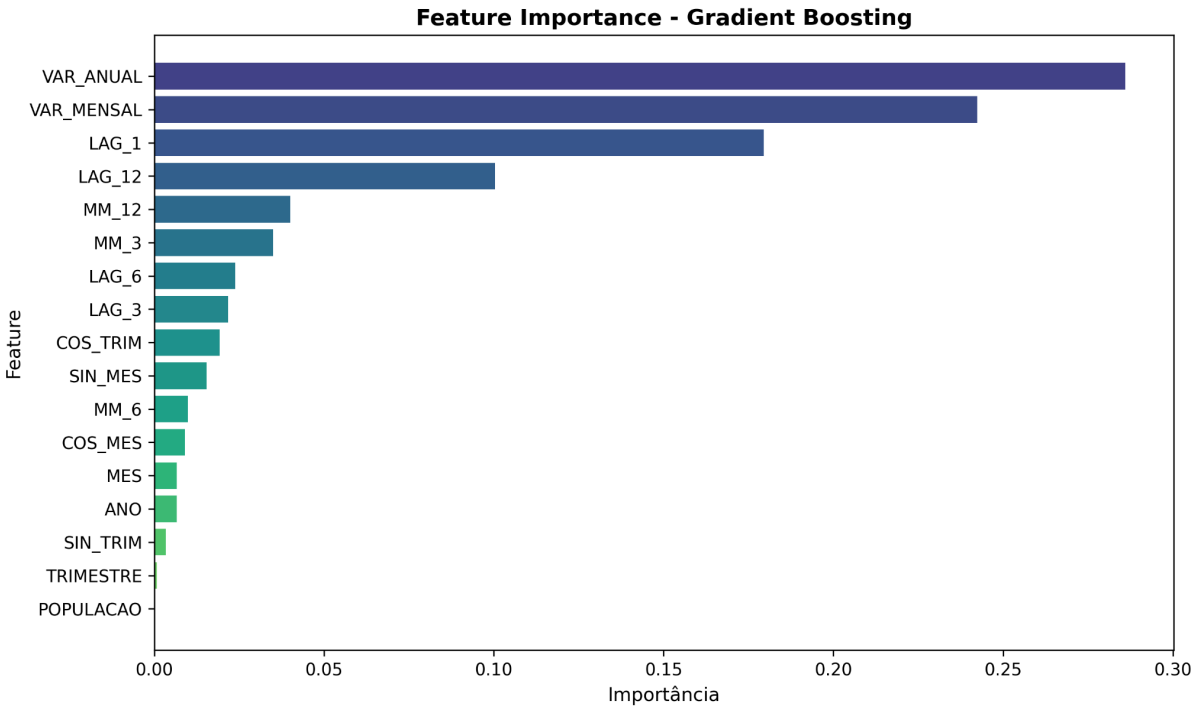

**Fonte:** Autores**.**

Esses resultados reforçam a importância de avaliar os modelos não apenas por métricas tradicionais de desempenho, mas também pela confiabilidade associada às previsões, especialmente em aplicações voltadas ao apoio à decisão em saúde pública.

### 5.2.1 Análise de Importância das Variáveis

A fim de compreender quais atributos exerceram maior influência na classificação do risco epidemiológico, foi realizada a análise de importância das variáveis com base no modelo final selecionado (Gradient Boosting Classifier). A importância foi calculada a partir da redução média do erro proporcionada por cada variável ao longo das árvores que compõem o ensemble.

Figura 10 – Gráfico de Importância de Features




**Fonte:** Autores.

Observa-se, na Figura 10, que as variáveis relacionadas à variação temporal apresentaram maior relevância para o modelo. A feature VAR_ANUAL (28,6%) foi a mais importante, seguida por VAR_MENSAL (24,2%), LAG_1 (17,9%) e LAG_12 (10,0%). Em conjunto, essas quatro variáveis representam 70,8% da importância total do modelo.

Esse resultado indica que o modelo baseou suas decisões principalmente em padrões de crescimento recente e comparações sazonais interanuais, reforçando a hipótese de que mudanças abruptas na incidência e persistência temporal são fatores determinantes para a classificação de surtos.

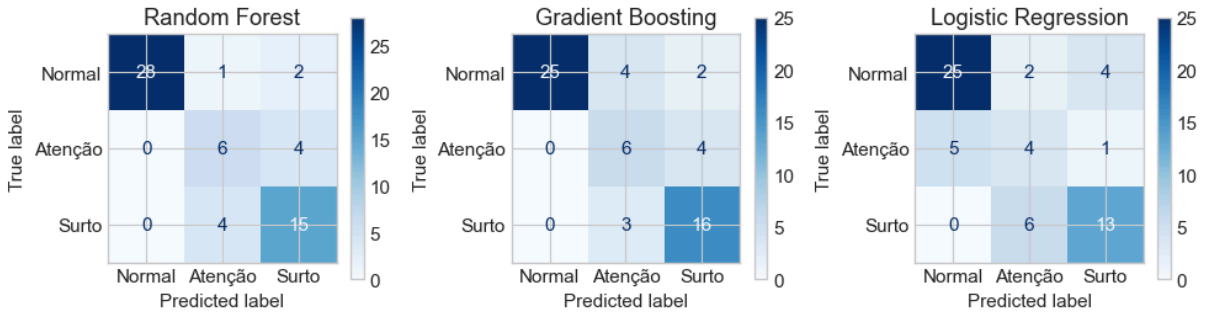
## 5.3 MATRIZES DE CONFUSÃO E ANÁLISE DE ERROS

Diferentemente das métricas agregadas, a análise das matrizes de confusão constitui uma etapa fundamental para a compreensão detalhada do comportamento dos modelos de classificação avaliados. Por meio disso, permite-se identificar padrões específicos de acerto e erro entre as classes de risco,

evidenciando como cada modelo se comporta na distinção entre estados epidemiológicos de Normalidade, Atenção e Surto.

Essa análise é particularmente relevante no contexto da vigilância em saúde pública, uma vez que diferentes tipos de erro possuem implicações distintas, especialmente aqueles associados à subestimação de cenários críticos. Assim, a avaliação das matrizes de confusão possibilita discutir não apenas o desempenho global dos modelos, mas também sua adequação operacional para a identificação precoce de surtos epidemiológicos.

Figura 11 – Matrizes de Confusão: *Random Forest* / *Gradient Boosting* / *Logistic Regression*




**Fonte:** Autores**.**

As matrizes de confusão evidenciam diferenças relevantes no comportamento dos modelos em relação às classes Normal, Atenção e Surto. O Random Forest apresentou bom desempenho na identificação da classe Normal, porém apresentou confusões recorrentes entre as classes Atenção e Surto, o que contribui para o aumento da incerteza observada.

O *Gradient Boosting* demonstrou maior consistência na identificação da classe Surto, com menor número de falsos negativos, aspecto crucial para evitar a subestimação de cenários epidemiológicos críticos. Embora ainda ocorram confusões entre Atenção e Surto, o modelo apresentou melhor equilíbrio entre sensibilidade e precisão para essa classe.

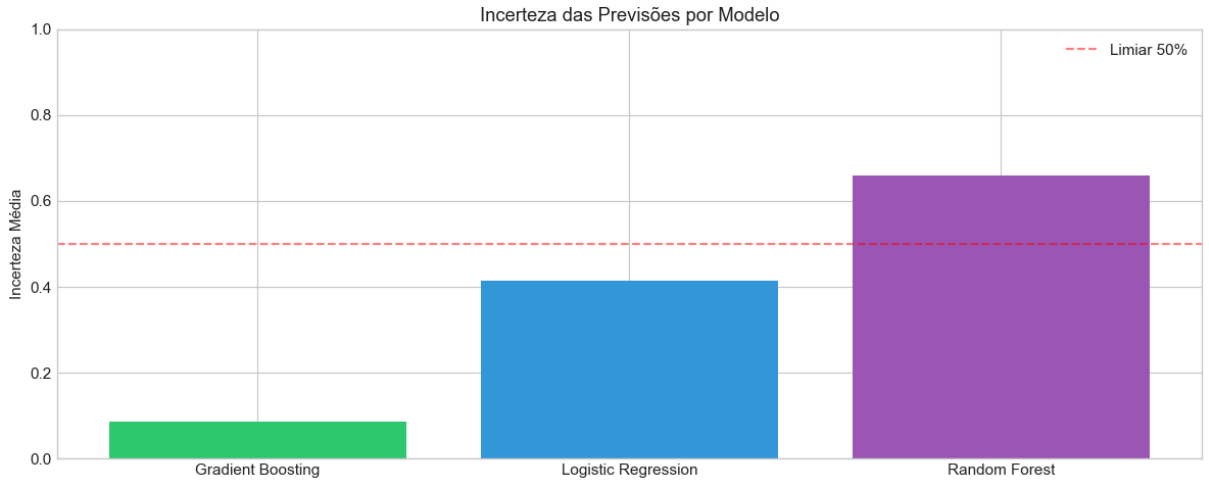
A Regressão Logística apresentou maior dificuldade em distinguir as três classes, com maior dispersão das previsões e confusões frequentes, especialmente entre Normal e Atenção. Esse comportamento sugere que modelos lineares podem

ser insuficientes para capturar a dinâmica não linear presente nos dados epidemiológicos analisados.

## 5.4 DESEMPENHO E INCERTEZA DAS PREVISÕES

O comparativo gráfico (Figura 12) das métricas reforça a superioridade do *Gradient Boosting* em termos de recall da classe Surto, enquanto o *Random Forest* apresenta ligeira vantagem em acurácia global. Entretanto, a análise da incerteza média evidencia um aspecto central desta pesquisa: desempenho elevado não implica necessariamente previsões confiáveis.

Figura 12 – Incerteza das Previsões por Modelo




**Fonte:** Autores.

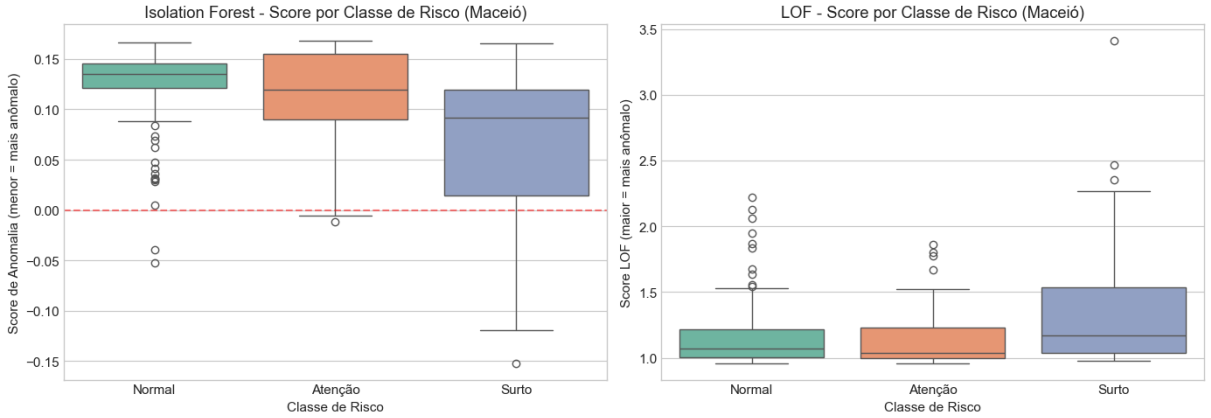
O *Random Forest*, apesar do bom desempenho médio, apresentou incerteza superior ao limiar de 50%, indicando que grande parte de suas previsões ocorre em regiões de baixa confiança. Em contraste, o *Gradient Boosting* manteve a incerteza média significativamente abaixo desse limiar, tornando-se mais adequado para aplicações em que a interpretação das previsões é tão importante quanto sua precisão.

Esses resultados validam a decisão metodológica de incorporar a incerteza como critério adicional de avaliação, alinhando-se às necessidades práticas da vigilância epidemiológica.

## 5.5 ANÁLISE DE ANOMALIAS E RELAÇÃO COM CLASSES DE RISCO

Os resultados da detecção de anomalias por meio dos algoritmos *Isolation Forest* e *Local Outlier Factor* (LOF) revelam padrões distintos entre as classes de risco epidemiológico. A Figura 13 apresenta, para cada classe (Normal, Atenção e Surto), a distribuição dos scores de anomalia obtidos por ambos os métodos. No caso do *Isolation Forest*, valores menores indicam maior grau de anomalia, enquanto no LOF valores mais elevados correspondem a maior atipicidade. Cada boxplot representa a mediana, os quartis e os valores extremos dos scores dentro de cada classe, permitindo comparar a dispersão e a concentração de observações atípicas entre os grupos.

Figura 13 – *Isolation Forest* e *Local Outlier Factor*: Score por Classe de Risco (Maceió - AL)




**Fonte:** Autores**.**

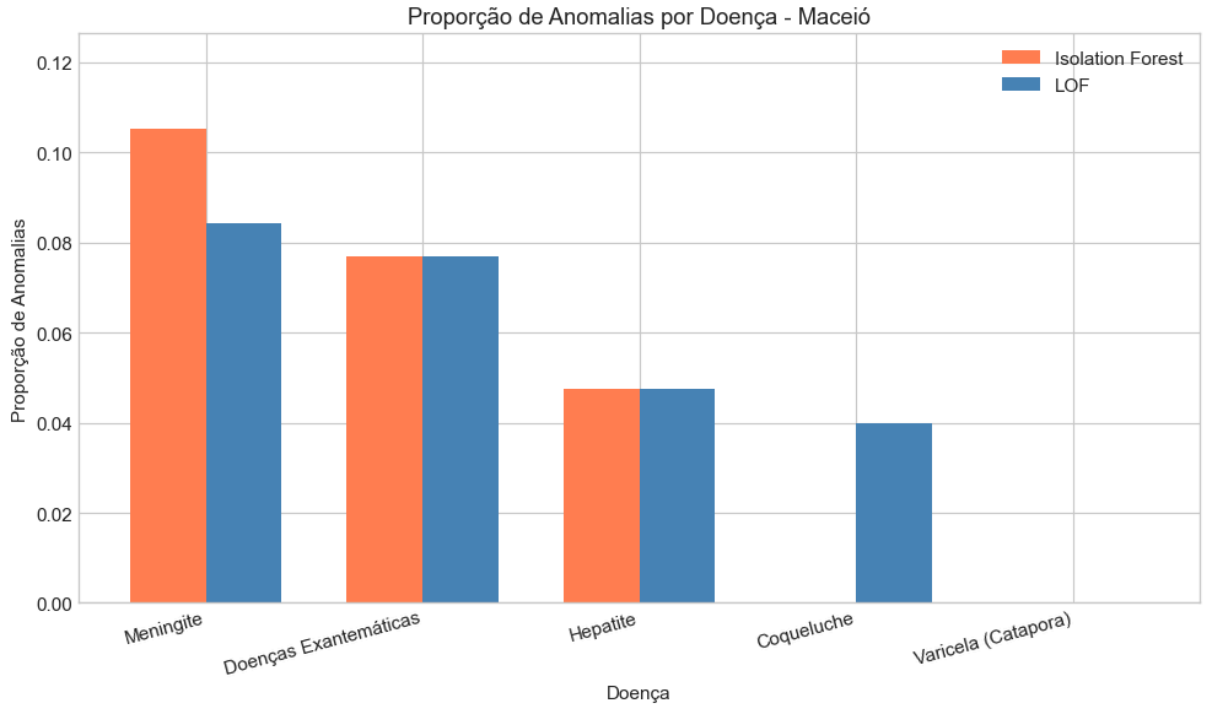
Observa-se que a classe Surto tende a apresentar maior variabilidade e valores mais extremos de score, especialmente no LOF, sugerindo que períodos classificados como surto frequentemente coincidem com padrões menos densos ou menos esperados na estrutura dos dados. Entretanto, a sobreposição entre as distribuições indica que não há separação absoluta entre as classes sob a ótica da anomalia.

A análise também demonstra que nem toda anomalia corresponde necessariamente a um surto epidemiológico. Registros classificados como Normal ou Atenção também apresentam valores elevados de anomalia, evidenciando que surtos e anomalias representam conceitos relacionados, porém distintos. Enquanto a classificação de surto baseia-se em limiares estatísticos históricos (média e desvio

padrão), a detecção de anomalias considera a estrutura global dos dados no espaço de atributos e sua densidade relativa. Essa constatação reforça a decisão de utilizar a detecção de anomalias como uma camada complementar de interpretação, e não como substituta da classificação de risco.

A proporção de anomalias identificadas por doença (Figura 14) permite observar diferenças na variabilidade temporal entre os agravos analisados. Doenças como meningite e doenças exantemáticas apresentam maior incidência de padrões atípicos, sugerindo comportamento episódico com flutuações mais abruptas quando comparadas a outras séries mais estáveis.

Figura 14 – Proporção de Anomalias por Doença em Maceió - AL



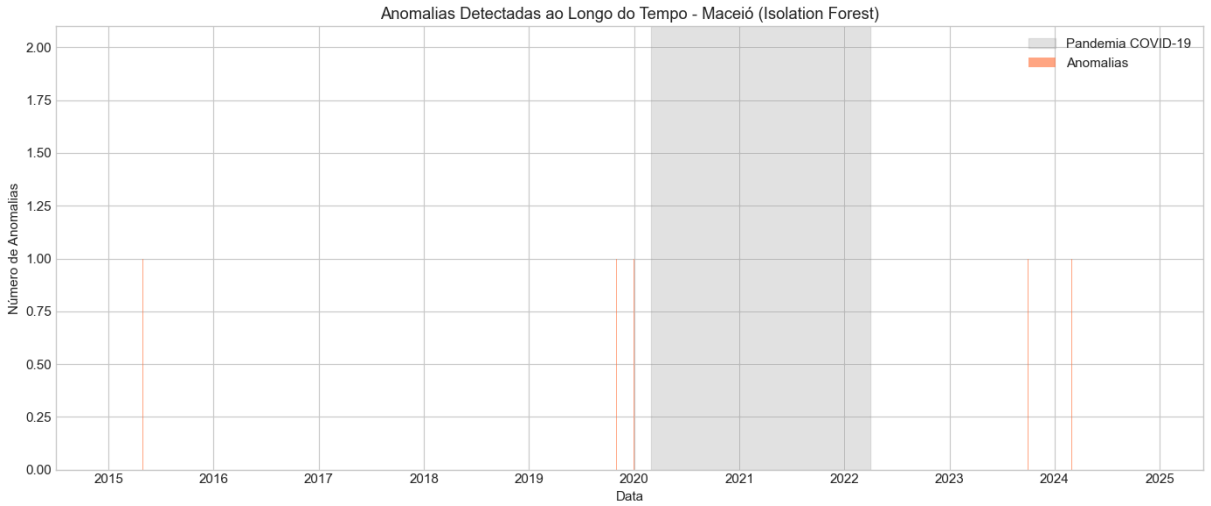

**Fonte:** Autores.

Nesse sentido, os resultados indicam que a detecção de anomalias fornece uma dimensão adicional de análise ao processo de vigilância epidemiológica. Embora exista associação entre surtos e padrões atípicos, a correspondência não é absoluta, o que evidencia a importância de combinar critérios estatísticos históricos com métodos baseados em estrutura de dados para interpretação mais robusta dos cenários epidemiológicos.

## 5.6 ANÁLISE TEMPORAL DAS ANOMALIAS

A análise temporal das anomalias detectadas (Figura 15) evidencia picos concentrados em períodos específicos, com destaque para o intervalo correspondente à pandemia de COVID-19. Esse comportamento justifica a exclusão desse período do treinamento dos modelos, uma vez que a pandemia introduziu alterações estruturais no processo de notificação e na dinâmica de transmissão de outras doenças infecciosas.

Figura 15 – Anomalias Detectadas ao Longo do Tempo em Maceió - AL com *Isolation Forest*




**Fonte:** Autores**.**

Mesmo após esse intervalo, observam-se eventos pontuais de anomalia em anos recentes, indicando que o sistema é capaz de identificar comportamentos incomuns fora de cenários extremos, o que reforça sua utilidade como ferramenta de apoio à vigilância contínua.

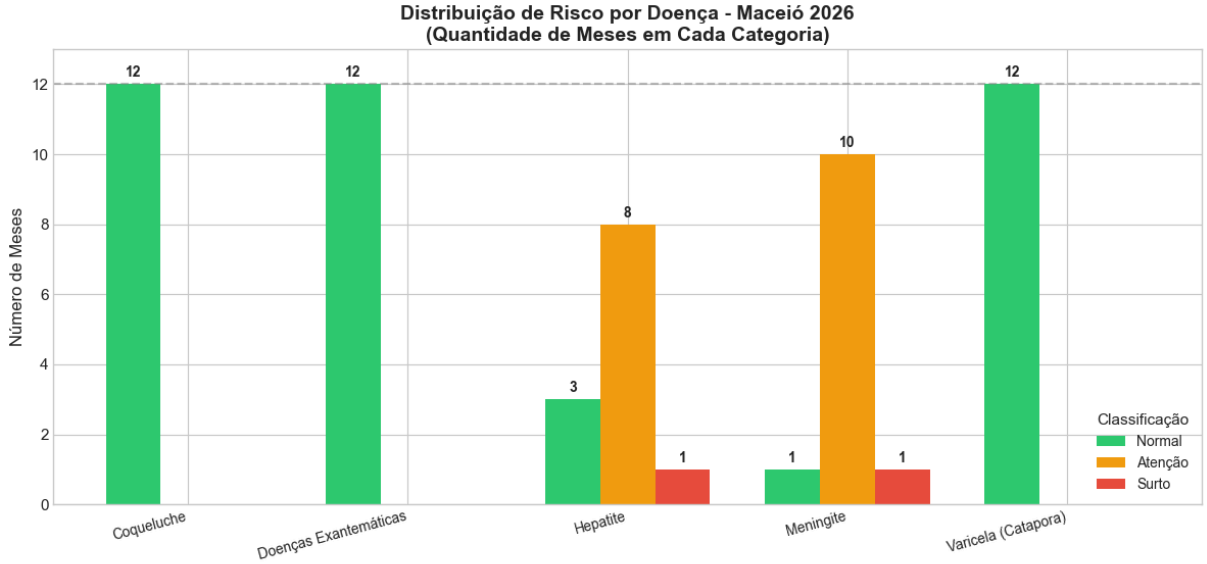
## 5.7 PREVISÕES DE RISCO EPIDEMIOLÓGICO PARA 2026

Diante do exposto, é apresentado nesta seção as previsões de risco epidemiológico geradas pelos modelos para o ano de 2026, considerando diferentes doenças analisadas no município de Maceió - AL. As previsões são discutidas sob múltiplas perspectivas, incluindo a distribuição global das classes de risco ao longo do ano, a evolução temporal mensal por doença e a síntese visual das estimativas, de modo a evidenciar tanto padrões gerais quanto comportamentos específicos relevantes para a vigilância em saúde.

### 5.7.1 Distribuição Global do Risco por Doença

A Figura 16 apresenta a distribuição do número de meses classificados em cada categoria de risco para as doenças analisadas ao longo de 2026. Observa-se que Coqueluche, Doenças Exantemáticas e Varicela (Catapora) foram classificadas como Normal durante todos os meses do ano, indicando estabilidade epidemiológica e baixo risco de ocorrência de surtos segundo o modelo preditivo.

Figura 16 – Distribuição de Risco por Doença




**Fonte:** Autores.

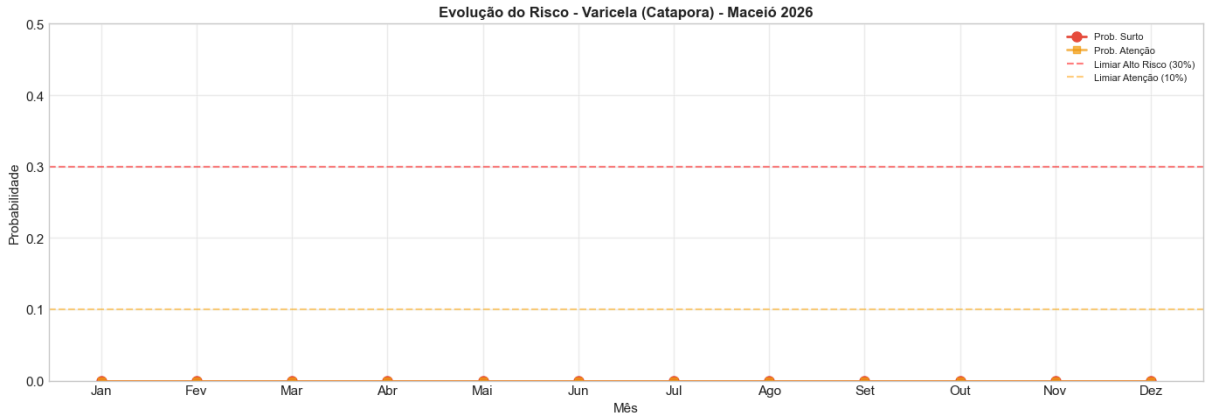
Em contrapartida, Hepatite e Meningite apresentaram maior variabilidade na classificação de risco. Para Hepatite, predominam meses classificados como Atenção, com a ocorrência pontual de um mês classificado como Surto, sugerindo um cenário de vigilância contínua ao longo do ano. Já a Meningite apresenta um número ainda maior de meses classificados como Atenção, além de um episódio de Surto, evidenciando um padrão mais instável quando comparado às demais doenças.

Esses resultados indicam que, embora a maioria dos agravos analisados apresente comportamento epidemiológico estável, determinadas doenças demandam monitoramento mais próximo, reforçando a importância de análises diferenciadas por agravo na vigilância epidemiológica.

### 5.7.2 Evolução Temporal do Risco por Doença

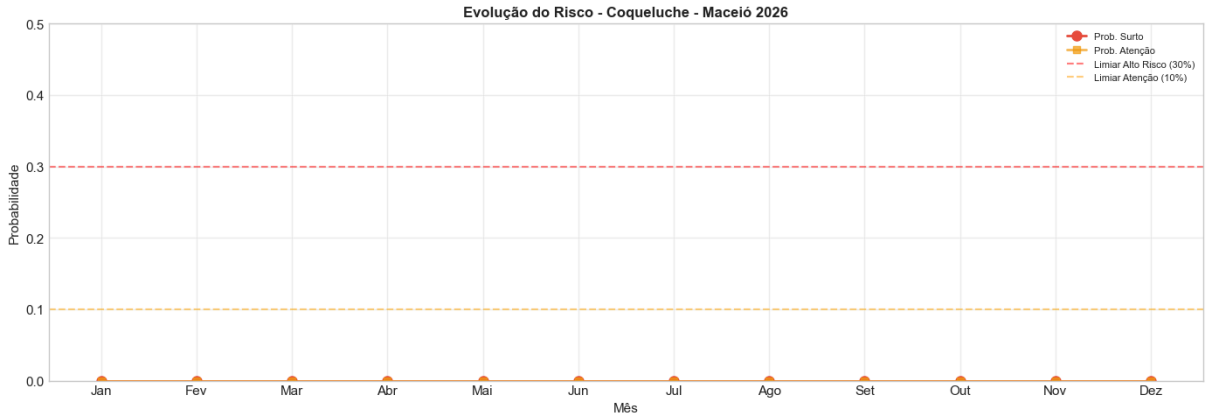
A análise da evolução temporal do risco permite observar, de forma detalhada, como as probabilidades associadas às classes Normal, Atenção e Surto se distribuem ao longo dos meses de 2026 para cada doença analisada.

Figura 17 – Evolução: Varicela



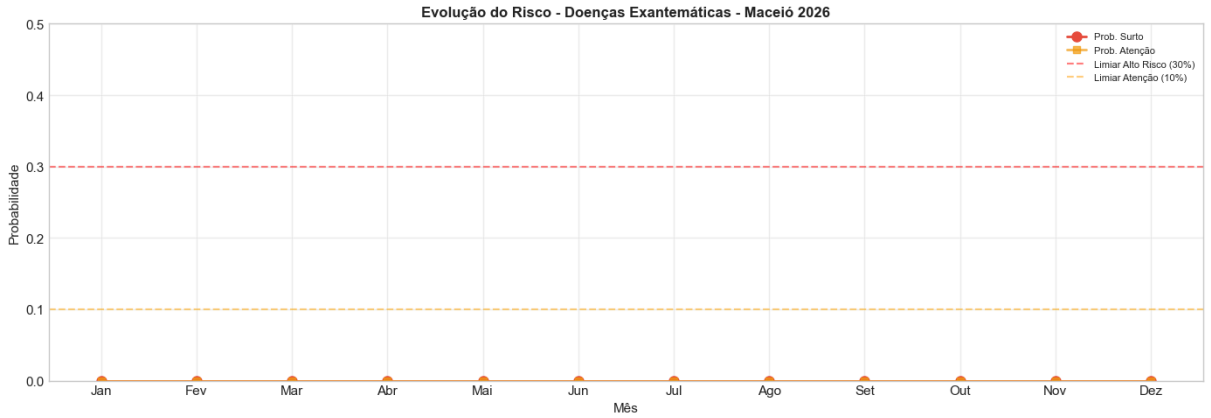

**Fonte:** Autores.

Figura 18 – Evolução: Coqueluche




**Fonte:** Autores.

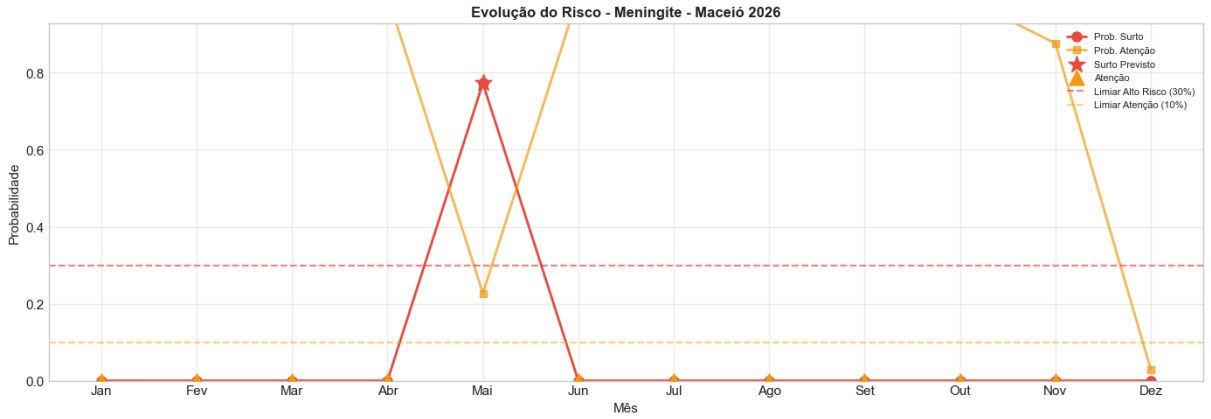
Figura 19 – Evolução: Doenças Exantemáticas




**Fonte:** Autores**.**

Para Varicela (Catapora), Coqueluche e Doenças Exantemáticas, as probabilidades de Surto permaneceram consistentemente abaixo do limiar de atenção ao longo de todo o período analisado, conforme evidenciado nas Figuras 17, 18 e 19, respectivamente. Esse comportamento reforça a classificação predominante como Normal, indicando estabilidade temporal e ausência de sinais de elevação significativa do risco epidemiológico.

Figura 20 – Evolução: Meningite

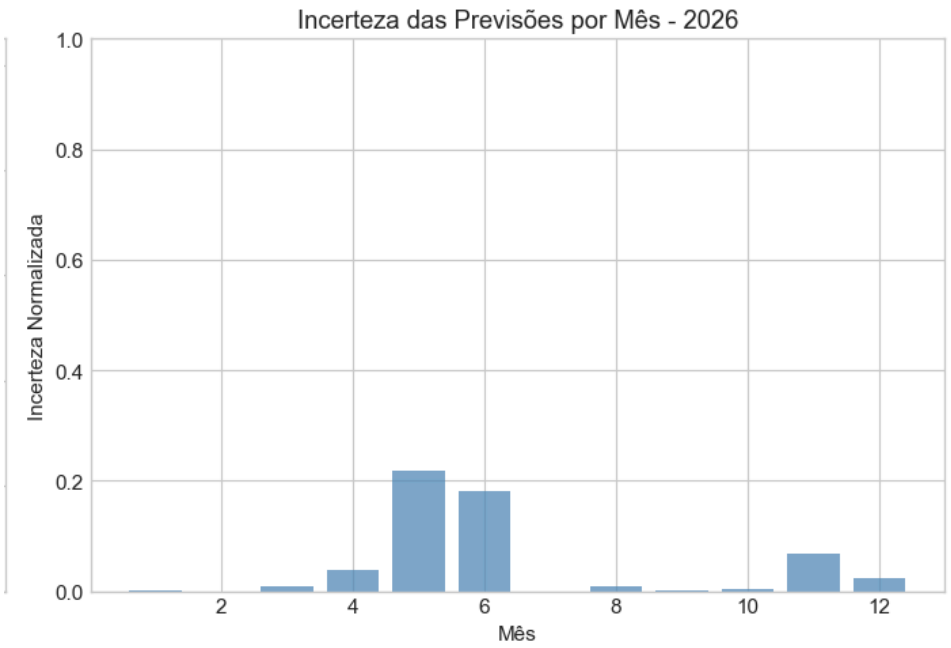



**Fonte:** Autores**.**

Em contraste, como é possível observar na Figura 20, Meningite apresenta um comportamento mais dinâmico, com elevação acentuada da probabilidade de Surto no mês de maio, ultrapassando o limiar de alto risco definido no modelo. Esse pico é acompanhado por períodos de Atenção em meses adjacentes, sugerindo um padrão de transição que pode preceder ou suceder eventos críticos. Tal comportamento evidencia a capacidade do modelo em capturar

variações abruptas e identificar janelas temporais potencialmente sensíveis para a vigilância.

Figura 21 – Gráfico de Incerteza das previsões ao longo dos meses de 2026

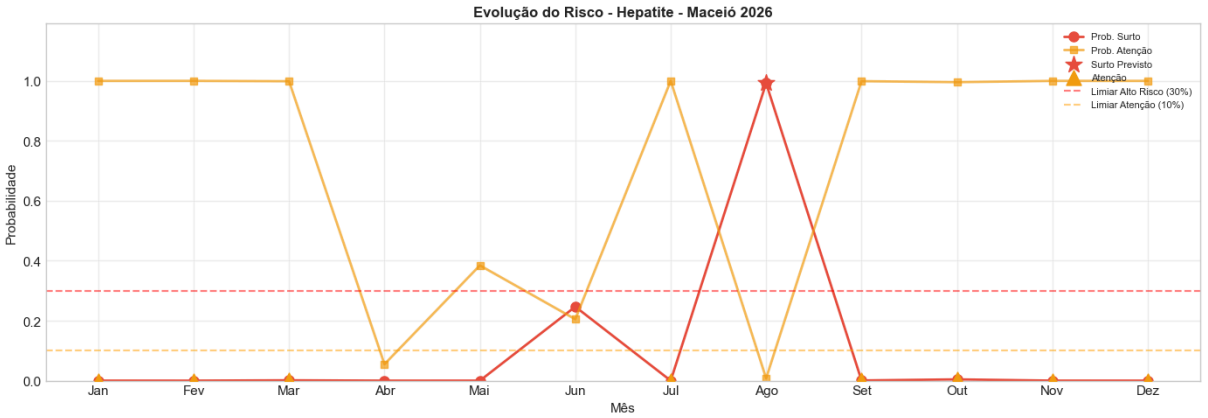



**Fonte:** Autores**.**

Todavia, apesar da classificação do mês de maio como Surto para hepatite, é necessário interpretar esse resultado com cautela. Conforme evidenciado na Figura 21, o mês de maio apresenta o maior valor de incerteza normalizada entre todos os meses analisados em 2026. Esse comportamento indica que, embora a probabilidade atribuída à classe Surto seja elevada, a distribuição das probabilidades entre as classes não é suficientemente concentrada, refletindo maior ambiguidade nos padrões históricos associados a esse período.

Assim, o surto identificado para hepatite em maio pode não representar um evento consolidado, mas sim um cenário limítrofe entre as classes Atenção e Surto, no qual pequenas variações nos dados podem influenciar significativamente a classificação final. Essa constatação reforça a importância de considerar simultaneamente a classe prevista e o grau de incerteza associado, especialmente em meses críticos, evitando interpretações categóricas baseadas exclusivamente na classificação de risco.

Figura 22 – Evolução: Hepatite




**Fonte:** Autores**.**

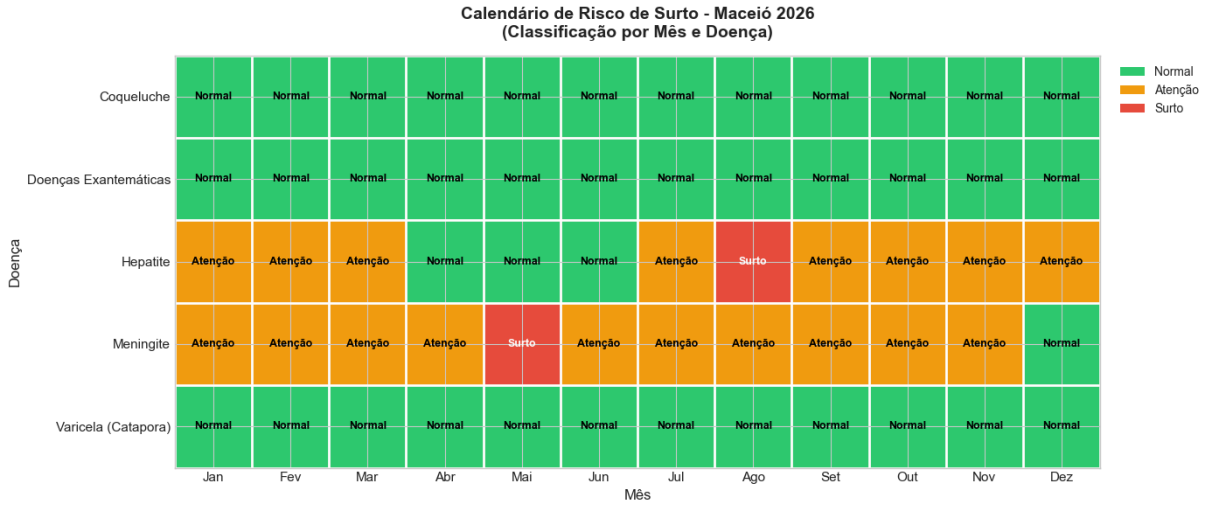
De forma semelhante, Hepatite (Figura 22) apresenta flutuações relevantes ao longo do ano, com predominância da classe Atenção e um pico expressivo de probabilidade de Surto no mês de agosto. Esse padrão indica um risco persistente, ainda que intermitente, reforçando a necessidade de acompanhamento contínuo ao longo do período analisado, mesmo nos meses em que o risco não atinge o nível máximo.

Essa análise temporal evidencia que a classificação de risco não se distribui de forma homogênea ao longo do ano, permitindo identificar meses críticos e períodos de maior instabilidade epidemiológica, aspecto fundamental para o planejamento de ações preventivas.

### 5.7.3 Síntese Visual e Interpretação Operacional do Risco

A síntese das previsões é apresentada por meio do calendário de risco epidemiológico, que consolida a classificação mensal por doença ao longo de 2026. Essa representação visual facilita a identificação de padrões recorrentes, períodos críticos e diferenças no comportamento epidemiológico entre os agravos analisados.

Figura 23 – Calendário de Risco para 2026

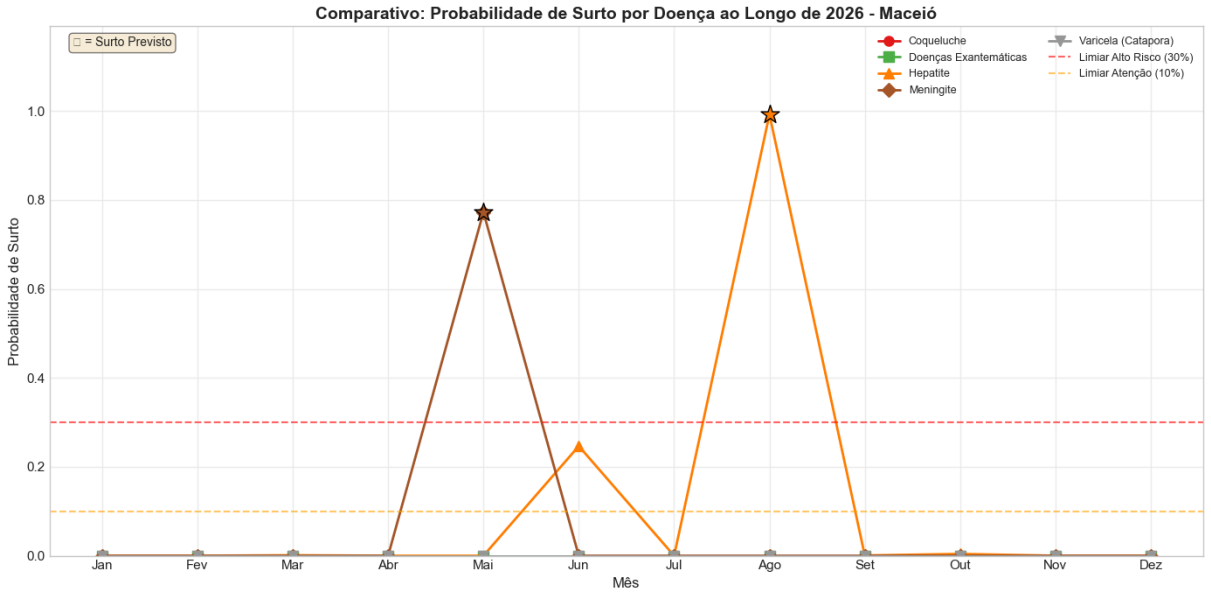



**Fonte:** Autores**.**

Observa-se na Figura 23 que os meses classificados como Surto concentram-se em pontos específicos do calendário, especialmente para Meningite e Hepatite, enquanto as demais doenças mantêm classificações estáveis ao longo do ano. A predominância da classe Atenção em determinados períodos reforça a utilidade da abordagem proposta, ao permitir ações graduais e proporcionais ao nível de risco estimado, em vez de respostas binárias baseadas apenas na ocorrência ou não de surtos.

Adicionalmente, a análise das probabilidades associadas às previsões evidencia que os eventos classificados como Surto apresentam valores significativamente superiores ao limiar de alto risco, enquanto os meses classificados como Atenção situam-se em faixas intermediárias de probabilidade. Essa separação contribui para maior interpretação dos resultados e fortalece o uso das previsões como subsídio à tomada de decisão em saúde pública.

Figura 24 – Comparativo Probabilidade de Surto ao longo de 2026




**Fonte:** Autores**.**

Em conjunto, os resultados obtidos para 2026, na Figura 24, demonstram que a solução proposta é capaz de fornecer previsões temporais consistentes, interpretáveis e alinhadas às necessidades operacionais da vigilância epidemiológica, permitindo não apenas a identificação de possíveis surtos, mas também a priorização de esforços de monitoramento e intervenção ao longo do tempo.

### 5.7.4 Síntese dos Resultados e Implicações da Vigilância

A análise integrada dos resultados evidencia que a abordagem proposta, baseada na classificação do risco epidemiológico aliada à quantificação da incerteza das previsões, constitui uma alternativa às estratégias tradicionais de previsão numérica de casos. Ao longo das análises, observou-se que modelos com desempenho semelhante em métricas globais podem apresentar comportamentos significativamente distintos quando avaliados sob a perspectiva da confiabilidade das previsões, aspecto crítico no contexto da vigilância em saúde pública.

Os resultados demonstram que a formulação do problema como uma tarefa de classificação em níveis de risco permite maior estabilidade frente a flutuações de curto prazo e à variabilidade inerente aos dados epidemiológicos. A introdução da classe intermediária de Atenção mostrou-se particularmente relevante, ao capturar cenários de transição que antecedem ou sucedem eventos de maior

gravidade, possibilitando a adoção de ações graduais e proporcionais ao risco estimado, em contraste com abordagens binárias centradas exclusivamente na detecção de surtos.

A avaliação comparativa dos modelos evidenciou que o *Gradient Boosting* apresentou o melhor equilíbrio entre sensibilidade à classe Surto, estabilidade preditiva e baixa incerteza, configurando-se como a abordagem mais adequada ao problema analisado. Esse resultado reforça a importância de critérios de avaliação alinhados aos objetivos operacionais da vigilância, nos quais a identificação precoce de cenários críticos deve ser priorizada em relação à acurácia global.

Do ponto de vista operacional, as previsões geradas para 2026 demonstram o potencial da solução proposta para apoiar o planejamento e a priorização de ações em saúde pública. A identificação de meses e doenças com maior risco estimado, associada à análise do grau de incerteza, permite diferenciar cenários que demandam monitoramento intensivo daqueles que requerem apenas acompanhamento rotineiro. Esse nível de detalhamento contribui para uma alocação mais eficiente de recursos e para a redução de respostas reativas baseadas exclusivamente em dados retrospectivos.

Por fim, os resultados obtidos evidenciam que a integração entre modelos preditivos, análise de incerteza e detecção de padrões anômalos oferece uma base metodológica consistente para o fortalecimento da vigilância epidemiológica. Embora as previsões devam ser interpretadas à luz das limitações inerentes aos dados e aos modelos utilizados, a abordagem proposta amplia a capacidade de antecipação e qualificação da análise epidemiológica, configurando-se como um instrumento promissor de apoio à tomada de decisão em contextos reais.

## 6 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo prospectar soluções baseadas em modelos preditivos para doenças infecciosas no município de Maceió – AL, por meio de técnicas de modelagem computacional e aprendizado de máquina, buscando comparar diferentes abordagens e identificar aquelas mais adequadas às características dos dados epidemiológicos disponíveis. À luz dos resultados obtidos, pode-se afirmar que o objetivo geral foi alcançado, uma vez que foram implementados, treinados e avaliados diferentes modelos preditivos, permitindo análise comparativa consistente entre as abordagens investigadas e possibilitando a identificação de estratégias com potencial para auxiliar na previsão de surtos.

No que se refere aos objetivos específicos, a prospecção de modelos computacionais foi realizada por meio da análise de abordagens distintas, incluindo algoritmos de aprendizado supervisionado baseados em regressão logística e métodos baseados em árvores de decisão. A comparação entre os modelos foi conduzida com o uso de métricas apropriadas de avaliação preditiva, considerando desempenho global, capacidade de generalização e estabilidade das previsões. Os resultados indicaram que técnicas baseadas em árvores de decisão, em especial o *Gradient Boosting*, apresentaram desempenho superior ao equilibrar sensibilidade à classe de maior relevância epidemiológica e menor grau médio de incerteza.

Adicionalmente, investigou-se o impacto de diferentes estratégias de treinamento, validação temporal e parametrização, confirmando que a escolha adequada de hiperparâmetros e a preservação da estrutura cronológica dos dados são fatores determinantes para a qualidade dos modelos desenvolvidos. A incorporação explícita da quantificação da incerteza das previsões constituiu um avanço metodológico, ampliando a interpretabilidade dos resultados e oferecendo suporte adicional à tomada de decisão.

Do ponto de vista aplicado, a solução proposta demonstrou potencial para apoiar processos de monitoramento e planejamento em saúde pública, permitindo a identificação prospectiva de cenários classificados como Normal, Atenção e Surto. Ao associar classificação de risco à estimativa de incerteza, o estudo contribui para uma abordagem mais cautelosa e informada, reduzindo a dependência exclusiva de análises retrospectivas e ampliando a capacidade de antecipação frente a possíveis eventos epidemiológicos.

## 6.1 LIMITAÇÕES E IMPEDIMENTOS

O desenvolvimento deste trabalho esteve condicionado a um conjunto de limitações de ordem metodológica, computacional e relacionada à disponibilidade e à qualidade dos dados, as quais influenciaram diretamente as decisões de escopo e delineamento da pesquisa.

Em um primeiro momento, buscou-se restringir a análise ao estado de Alagoas, com o intuito de reduzir o volume de dados e de viabilizar o processamento. Ainda assim, persistiram limitações relevantes, especialmente relacionadas à indisponibilidade de séries históricas completas do PNI anteriores ao período da pandemia de COVID-19. Enquanto os dados do SINAN utilizados neste estudo abrangem aproximadamente os últimos dez anos, os registros de imunização disponíveis apresentavam cobertura consistente apenas a partir de 2020, inviabilizando análises temporais integradas e comparáveis entre os dois sistemas.

Assim, observou-se uma distribuição fortemente desigual do número de registros entre os diferentes agravos analisados, com doenças como hepatite e meningite apresentando volumes expressivamente superiores, enquanto outros agravos, como a raiva, registravam números extremamente reduzidos, por vezes limitados a poucos casos por ano. Essa assimetria resultou em um viés potencial no processo de treinamento dos modelos, favorecendo doenças com maior frequência de registros e dificultando a generalização para agravos raros.

Um dos principais fatores limitantes identificados foi o impacto da pandemia de COVID-19 sobre a qualidade e a completude dos dados epidemiológicos. Conforme discutido na fundamentação teórica, a pandemia introduziu alterações estruturais nos processos de vigilância e notificação, intensificando a subnotificação de diversos agravos.

Somado a isso, a aplicação da metodologia está condicionada à qualidade e à granularidade dos dados disponíveis, o que pode impactar na generalização para outros contextos ou escalas geográficas. Dessa forma, os resultados obtidos devem ser interpretados com cautela, especialmente no que se refere à sua replicabilidade em cenários com diferentes estruturas de dados ou níveis de maturidade dos sistemas de informação em saúde.

Apesar dessas limitações, as decisões adotadas permitiram a construção de um conjunto de dados mais coerente com os objetivos do estudo, reduzindo

vieses extremos e possibilitando a aplicação e a avaliação das técnicas de modelagem computacional propostas.

## 6.2 TRABALHOS FUTUROS

Como desdobramento natural desta pesquisa, diversos caminhos podem ser explorados com o objetivo de ampliar o escopo, a robustez metodológica e a aplicabilidade prática da abordagem proposta. Assim, como trabalhos futuros, destacam-se as seguintes direções:

Primeiramente, a possibilidade de ampliação e diversificação das fontes de dados utilizadas na modelagem. Uma extensão natural consiste em integrar informações do Programa Nacional de Imunizações (PNI), desde que acompanhadas de estratégias mais adequadas de ingestão incremental, particionamento e versionamento, de modo a contornar limitações de volume e disponibilidade identificadas neste trabalho. A incorporação de variáveis de cobertura vacinal, atraso vacinal e variações sazonais de imunização pode contribuir para aumentar o poder explicativo dos modelos e permitir análises mais completas sobre fatores associados ao risco de surtos.

Adicionalmente, sugere-se a avaliação do pipeline e dos modelos em diferentes granularidades temporais. Neste estudo, os dados foram agregados mensalmente; contudo, a vigilância epidemiológica opera de forma relevante também em semanas epidemiológicas. A migração para séries semanais pode aumentar a capacidade de antecipação e tornar as previsões mais acionáveis, além de permitir comparação com rotinas usuais de monitoramento.

Outra possibilidade potencialmente relevante consiste na validação externa da abordagem proposta, por meio de sua aplicação em outros contextos geográficos ou epidemiológicos. A aplicação do método em diferentes municípios ou estados permitiria avaliar sua capacidade de generalização e identificar a necessidade de ajustes para lidar com variações regionais nos padrões de notificação e transmissão das doenças.

Por fim, trabalhos futuros podem investigar a implementação operacional da solução em ambientes reais de vigilância, integrando os modelos a sistemas de monitoramento contínuo e painéis interativos. Essa etapa permitiria avaliar não apenas o desempenho técnico dos modelos, mas também sua utilidade prática, aceitabilidade e impacto no processo decisório de profissionais de saúde.

## REFERÊNCIAS

AGUIAR, Tamires Saraiva *et al*. Epidemiological profile of meningitis in Brazil, based on data from DataSUS in the years 2020 and 2021. ***Research, Society and Development***, [S. l.], v. 11, n. 3, p. e50811327016, 2022. DOI: 10.33448/rsd-v11i3.27016. Disponível em: https://rsdjournal.org/rsd/article/view/27016. Acesso em: 3 fev. 2026.

AHMAD, Ghulab Nabi *et al*. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV. ***Ieee Access***, v. 10, p. 80151-80173, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3165792.

ARMBRUST, Michael et al. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In: ***Proceedings of CIDR***. 2021. p. 28. Disponível em: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf. Acesso em: 3 fev. 2026.

ANTUNES, José Leopoldo Ferreira; CARDOSO, Maria Regina Alves. Uso da análise de séries temporais em estudos epidemiológicos. ***Epidemiologia e Serviços de Saúde***, Brasília, v. 24, n. 3, p. 565–576, 2015. Disponível em: https://www.scielo.br/j/ress/a/ZV4h6JY7p9Yw5xXWvM3xjvN/. Acesso em: 22 fev. 2026.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Guia de Vigilância em Saúde. Brasília: Ministério da Saúde, 2022. Disponível em: https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/publicacoes-svs/vigilancia/guia-de-vigilancia-em-saude-5a-edicao. Acesso em: 22 fev. 2026.

BALCAN, D. *et al*. Multiscale mobility networks and the spatial spreading of infectious diseases. ***Proceedings of the National Academy of Sciences***, v. 106, n. 51, p. 21484–21489, 2009. DOI: https://doi.org/10.1073/pnas.0906910106.

BATISTA, Lucas R.; SILVA, Ana Paula. Qualidade de dados em pipelines analíticos: desafios e estratégias. ***Revista Brasileira de Sistemas de Informação***, v. 15, n. 2, p. 45-60, 2022. Disponível em: https://periodicos.ufpb.br/index.php/rbsi. Acesso em: 28 jan. 2026.

BOING, Antonio Fernando; FONSECA, Maria de Jesus Mendes da. Pesquisa em epidemiologia no Brasil: desafios, caminhos e compromissos para o futuro. ***Revista Brasileira de Epidemiologia***, v. 28, supl. 1, e250003supl1, 2025. DOI: 10.1590/1980-549720250003.supl.1. Disponível em: https://doi.org/10.1590/1980-549720250003.supl.1. Acesso em: 03 fev. 2026.

BORGES, Pollyanna Kassia de Oliveira et al. Impacto da COVID-19 sobre doenças de notificação compulsória: um estudo de série temporal. ***Revista da Escola de Enfermagem da USP***, v. 58, p. e20240098, 2024. DOI: https://doi.org/10.1590/1980-220X-REEUSP-2024-0098pt.

BREIMAN, Leo. Random forests. ***Machine Learning***, Dordrecht, v. 45, n. 1, p. 5–32, out. 2001. DOI: https://doi.org/10.1023/A:1010933404324.

BREUNIG, Markus M. et al. LOF: identifying density-based local outliers. In: ***Proceedings of the 2000 ACM SIGMOD international conference on Management of data***. 2000. p. 93-104. DOI: https://doi.org/10.1145/342009.335388.

BRITO, Mariana et al. Completeness of notifications of accidents involving venomous animals in the Information System for Notifiable Diseases: a descriptive study, Brazil, 2007-2019. ***Epidemiologia e Serviços de Saúde***, v. 32, p. e2022666, 2023. DOI: https://doi.org/10.1590/S2237-96222023000100002.

CABRERA, Maritza et al. Dengue prediction in Latin America using machine learning and the one health perspective: a literature review. ***Tropical Medicine and Infectious Disease***, v. 7, n. 10, p. 322, 2022. DOI: https://doi.org/10.3390/tropicalmed7100322.

CERRI, Ricardo; CARVALHO, André Carlos Ponce de Leon Ferreira de. Aprendizado de máquina: breve introdução e aplicações. ***Cadernos de Ciência & Tecnologia***, v. 34, n. 3, p. 297-313, 2017. DOI: https://doi.org/10.35977/0104-1096.cct2017.v34.26381

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. Anomaly detection: A survey. ***ACM computing surveys (CSUR)***, v. 41, n. 3, p. 1-58, 2009. DOI: https://doi.org/10.1145/1541880.1541882

DIXON, James. Pentaho, Hadoop, and Data Lakes. ***James Dixon's Blog***, 2010. Disponível em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/. Acesso em: 5 fev. 2026.

DU, Lin; PANG, Yan. A novel data-driven methodology for influenza outbreak detection and prediction. ***Scientific reports*****,** v. 11, n. 1, p. 13275, 2021. DOI: https://doi.org/10.1038/s41598-021-92484-6.

FERREIRA, Manuela Klanovicz; VANZ, Samile Andrea de Souza. Reprodutibilidade em e-science: uma visão geral dos conceitos relacionados e das ferramentas de suporte mais citadas. ***Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação***, v. 30, e103385, 2025. DOI: 10.5007/1518-2924.2025.e103385. Disponível em: https://doi.org/10.5007/1518-2924.2025.e103385. Acesso em: 03 fev. 2026.

FONTANA, Éliton. Introdução aos algoritmos de aprendizagem supervisionada. ***Departamento de Engenharia Química, Universidade Federal do Paraná***, 2020. Disponível em: <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>. Acesso em: 02 fev. 2026.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. ***Annals of statistics***, p. 1189-1232, 2001. Disponível em: http://www.jstor.org/stable/2699986. Acesso em: 13 fev. 2026.

GAO, Shan et al. Early detection of disease outbreaks and non-outbreaks using incidence data. ***arXiv preprint arXiv:2404.08893***, 2024. DOI: https://doi.org/10.48550/arXiv.2404.08893.

GOMES, Dennis dos Santos. Inteligência Artificial: conceitos e aplicações. ***Revista Olhar Científico***, v. 1, n. 2, p. 234-246, 2010. Disponível em: https://www.professores.uff.br/screspo/wp-content/uploads/sites/127/2017/09/ia_intro.pdf. Acesso em: 28 jan. 2026.

HASTIE, Trevor. The elements of statistical learning: data mining, inference, and prediction. 2009. DOI: https://doi.org/10.1111/j.1541-0420.2010.01516.x.

INMON, Bill. **Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump**. *Technics Publications*, LLC, 2016.

JONES, Kate E. et al. Global trends in emerging infectious diseases. ***Nature***, v. 451, n. 7181, p. 990-993, 2008. DOI: https://doi.org/10.1038/nature06536

KOIKE, Marcia. DataSUS: Uma Ferramenta Essencial para a Saúde Pública no Brasil. ***Arquivos Brasileiros de Cardiologia***, v. 122, n. 2, p. e20250123, 2025. DOI: https://doi.org/10.36660/abc.20250123.

LAMER, Antoine et al. Data lake, data warehouse, datamart, and feature store: Their contributions to the complete data reuse pipeline. ***JMIR medical informatics***, v. 12, p. e54590, 2024. DOI: https://doi.org/10.2196/54590.

LIMA, Emanuelle C. A.; OLIVEIRA, João P.; SANTOS, Mariana R. Qualidade da informação no Sistema de Informação de Agravos de Notificação: uma revisão integrativa. ***Cadernos de Saúde Pública***, Rio de Janeiro, v. 37, n. 6, e00123420, 2021. Disponível em: https://www.scielo.br/j/csp/. Acesso em: 28 jan. 2026.

LIU, Fei Tony; TING, Kai Ming; ZHOU, Zhi-Hua. Isolation forest. In: ***2008 eighth ieee international conference on data mining***. IEEE, 2008. p. 413-422. DOI: https://doi.org/10.1109/ICDM.2008.17.

LIU, Rui et al. Optimizing data pipelines for machine learning in feature stores. ***Proceedings of the VLDB Endowment***, v. 16, n. 13, p. 4230-4239, 2023. DOI: https://dl.acm.org/doi/10.14778/3625054.3625060.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. ***Estudos avançados***, v. 35, p. 85-94, 2021. DOI: https://doi.org/10.1590/s0103-4014.2021.35101.007.

MAIA, Daniely Aleixo Barbosa et al. Avaliação da implantação do Sistema de Informação de Agravos de Notificação em Pernambuco, 2014. ***Epidemiologia e Serviços de Saúde***, v. 28, p. e2018187, 2019. DOI: https://doi.org/10.5123/S1679-49742019000100002.

MENEZES, Renata PB de; SCOTTI, Luciana; SCOTTI, Marcus T. Aprendizado de máquina aplicado a qsar. ***Química Nova***, v. 47, n. 7, p. e-20240024, 2024. DOI: https://doi.org/10.21577/0100-4042.20240024.

NARGESIAN, Fatemeh *et al*. Data lake management: challenges and opportunities. ***Proceedings of the VLDB Endowment***, v. 12, n. 12, p. 1986-1989, 2019. DOI: https://doi.org/10.14778/3352063.3352116.

ORR, Laurel et al. Managing ml pipelines: feature stores and the coming wave of embedding ecosystems. ***arXiv preprint arXiv:2108.05053***, 2021. DOI: https://doi.org/10.48550/arXiv.2108.05053.

PAGOTTO, Daniel do Prado; MARQUES, Wanderson da Silva; OLIVEIRA, Denise Santos de; FERREIRA, Vicente da Rocha Soares; AZEVEDO, Vinicius Nunes de; BORGES JÚNIOR, Cândido Vieira. Inovação em saúde: a implementação de um data lake para armazenamento, sistematização e disponibilização de dados em saúde no Brasil. ***InCID: Revista de Ciência da Informação e Documentação***, v. 15, n. 1, 2024. DOI: 10.11606/issn.2178-2075.incid.2024.213345. Disponível em: https://doi.org/10.11606/issn.2178-2075.incid.2024.213345. Acesso em: 03 fev. 2026.

PAIXÃO, Gabriela Miana de Mattos et al. Machine learning na medicina: revisão e aplicabilidade. ***Arquivos Brasileiros de Cardiologia***, v. 118, n. 1, p. 95-102, 2022. DOI: https://doi.org/10.36660/abc.20200596.

ROCHA, M. S.; BARTHOLOMAY, P. et al. Notifiable Diseases Information System (SINAN): main characteristics of notification and data analysis related to tuberculosis. Epidemiologia e Serviços de Saúde, Brasília, v. 29, n. 1, e2019017, 2020. Disponível em: https://www.scielosp.org/article/ress/2020.v29n1/e2019017/en/. Acesso em: 3 fev. 2026.

SALLAS, Janaína et al. Decréscimo nas notificações compulsórias registradas pela Rede Nacional de Vigilância Epidemiológica Hospitalar do Brasil durante a pandemia da COVID-19: um estudo descritivo, 2017-2020. ***Epidemiologia e Serviços de Saúde***, v. 31, p. e2021303, 2022. DOI: https://doi.org/10.1590/S1679-49742022000100011.

SANTANGELO, O. E. *et al*. Machine learning and prediction of infectious diseases: a systematic review. ***Machine Learning and Knowledge Extraction***, v. 5, n. 1, p. 175–198, 2023. DOI: https://doi.org/10.3390/make5010013.

SANTOS, Ricardo M.; COSTA, João V. Pipelines de dados e arquitetura para ciência de dados. ***Revista Tecnologia da Informação e Comunicação***, v. 10, n. 1, p. 22-35, 2020. Disponível em: https://revistas.ufpr.br/rtic. Acesso em: 28 jan. 2026.

SCHLEDER, Gabriel R.; FAZZIO, Adalberto. Machine Learning na Física, Química, e Ciência de Materiais: Descoberta e Design de Materiais. ***Revista Brasileira de Ensino de Física***, v. 43, p. e20200407, 2021. DOI: 10.1590/1806-9126-RBEF-2020-0407.

SCURSONE, Gabriel Fuscald *et al*. Hyperparameter Optimization of XGBoost on Air Pollution and Respiratory Health Data. ***Studies in Health Sciences***, v. 6, n. 4, p. e21945-e21945, 2025. DOI: https://doi.org/10.54022/shsv6n4-035.

SHAMAN, Jeffrey; KARSPECK, Alicia. Forecasting seasonal outbreaks of influenza. ***Proceedings of the National Academy of Sciences***, v. 109, n. 50, p. 20425-20430, 2012. DOI: https://doi.org/10.1073/pnas.1208772109.

SILVA, Gabriela Drummond Marques da et al. Identificação de microrregiões com subnotificação de casos de tuberculose no Brasil, 2012 a 2014. ***Epidemiologia e***

***Serviços de Saúde***, v. 29, p. e2018485, 2020. DOI: https://doi.org/10.5123/S1679-49742020000100025.

SILVA, Marcos A.; OLIVEIRA, Fernanda L.; PEREIRA, Tiago S. Arquiteturas de dados para análise e aprendizado de máquina. ***Revista Brasileira de Computação Aplicada***, v. 13, n. 3, p. 85-98, 2021. Disponível em: https://seer.upf.br/index.php/rbca. Acesso em: 28 jan. 2026.

SOUZA, Larissa M.; SILVA, Carlos A. Uso do SINAN como ferramenta para a vigilância epidemiológica no Brasil. ***Revista Brasileira de Epidemiologia***, São Paulo, v. 23, e200045, 2020. Disponível em: https://www.scielo.br/j/rbepid/. Acesso em: 28 jan. 2026.

WANG, Shujuan *et a*l. Research on expansion and classification of imbalanced data based on SMOTE algorithm. ***Scientific reports***, v. 11, n. 1, p. 24039, 2021. DOI: https://doi.org/10.1038/s41598-021-03430-5.

ZHANG, Yiming et al. An intelligent early warning system of analyzing Twitter data using machine learning on COVID-19 surveillance in the US. ***Expert Systems with Applications***, v. 198, p. 116882, 2022. DOI: https://doi.org/10.1016/j.eswa.2022.116882

ZANARDO, Giovanni E. *et al*. Uma comparação entre métodos baseados em aprendizado de máquina para inferir número de casos semanais de dengue. In: ***Seminário Integrado de Software e Hardware (SEMISH)***. SBC, 2024. p. 37-48. DOI: https://doi.org/10.5753/semish.2024.1921.