CEFET/RJ - CENTRO FEDERAL DE EDUCACAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA

Modelagem Preditiva de Internações Respiratórias a partir de Poluentes Atmosféricos e Variáveis Meteorológicas

Bianca Gallicchio Tavares

João Henrique dos Reis Terêncio

Prof. Orientador:

Eduardo Bezerra, D.Sc.

Orientador

Rio de Janeiro,

6 de Janeiro de 2026

CEFET/RJ - CENTRO FEDERAL DE EDUCACAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA

Modelagem Preditiva de Internações Respiratórias a partir de Poluentes Atmosféricos e Variáveis Meteorológicas

Bianca Gallicchio Tavares

João Henrique dos Reis Terêncio

Projeto final apresentado em cumprimento às normas do Departamento de Educação Superior do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos para obtenção do título de Bacharel em Ciência da Computação.

Prof. Orientador:

Eduardo Bezerra, D.Sc.

Orientador

Rio de Janeiro,

6 de Janeiro de 2026

Obtenha a ficha catalográfica junto a biblioteca.
Substitua o arquivo ficha.pdf pela versão obtida lá.

DEDICATÓRIA

Dedicamos este trabalho aos nossos pais, cujo apoio, carinho e dedicação foram essenciais em toda a nossa trajetória acadêmica. Sem os sacrifícios e a estrutura oferecida por eles, este trabalho não teria sido possível. Dedicamos também aos nossos familiares e amigos, pelo incentivo e pela compreensão ao longo do caminho, contribuindo para a construção do nosso aprendizado.

AGRADECIMENTOS

Agradecemos à CAPES, ao CNPq e à FAPERJ pelo apoio e financiamento parcial desta pesquisa. Agradecemos ao CEFET/RJ pela formação acadêmica e pela infraestrutura disponibilizada ao longo do curso. Agradecemos também à professora Rejane Sobrino Pinheiro pelas sugestões de melhorias e pela disponibilidade em esclarecer dúvidas ao longo do desenvolvimento deste trabalho. Por fim, agradecemos, em especial, ao professor e orientador Eduardo Bezerra pela orientação, disponibilidade e pelos conhecimentos compartilhados, fundamentais para a realização desta pesquisa.

RESUMO

A relação entre poluição atmosférica, condições meteorológicas e internações por doenças respiratórias tem motivado o uso crescente de modelos de aprendizado de máquina para previsão de desfechos em saúde pública. Entretanto, apesar do avanço metodológico, a literatura ainda apresenta lacunas relacionadas à heterogeneidade das variáveis utilizadas, limitações de generalização em diferentes contextos urbanos e diversidade de abordagens preditivas. Este trabalho busca preencher esse espaço ao realizar uma revisão sistemática e estruturada da produção científica que utiliza algoritmos de aprendizado de máquina para prever internações e atendimentos respiratórios com base em poluentes atmosféricos e variáveis meteorológicas. O objetivo é identificar convergências metodológicas, lacunas existentes e direções relevantes para o desenvolvimento de modelos preditivos robustos aplicáveis ao contexto brasileiro. Os resultados da revisão sistemática mostram que modelos como Random Forest, MLP, SVR, LSTM e ensembles híbridos apresentam desempenho superior aos modelos lineares tradicionais, especialmente quando incorporam variáveis defasadas, meteorológicas e sazonais. Poluentes como PM2.5, PM10 e NO2 surgem de forma consistente como preditores relevantes, e fatores meteorológicos como temperatura, umidade e vento complementam significativamente o desempenho dos modelos. A literatura também destaca avanços em interpretabilidade, por meio de técnicas como SHAP e LIME, e evidencia que estudos realizados no Brasil enfrentam desafios de cobertura e qualidade dos dados, mas ainda assim apresentam resultados sólidos. As conclusões indicam que os métodos de aprendizado de máquina representam ferramentas promissoras para previsão de internações respiratórias, mas ainda demandam maior padronização metodológica, integração de múltiplas fontes de dados e desenvolvimento de modelos mais explicáveis e adaptáveis a diferentes realidades urbanas.

Palavras-chaves: Doenças respiratórias. Poluição atmosférica. Aprendizado de máquina. Modelagem preditiva. Saúde pública urbana.

ABSTRACT

The relationship between air pollution, meteorological conditions, and hospital admissions for respiratory diseases has driven the growing use of machine learning models to predict public health outcomes. However, despite methodological advances, the literature still presents gaps related to the heterogeneity of variables used, limitations in generalization across different urban contexts, and diversity of predictive approaches. This work seeks to fill this gap by conducting a systematic and structured review of scientific studies that employ machine learning algorithms to predict respiratory hospitalizations and outpatient visits based on air pollutants and meteorological variables. The goal is to identify methodological convergences, existing gaps, and relevant directions for developing robust predictive models applicable to the Brazilian context. The results of the systematic review show that models such as Random Forest, MLP, SVR, LSTM, and hybrid ensembles outperform traditional linear models, especially when incorporating lagged, meteorological, and seasonal variables. Pollutants such as PM2.5, PM10, and NO2 consistently emerge as relevant predictors, while meteorological factors such as temperature, humidity, and wind significantly enhance model performance. The literature also highlights advances in interpretability through techniques such as SHAP and LIME and shows that studies conducted in Brazil face challenges related to data coverage and quality, yet still present solid results. The conclusions indicate that machine learning methods represent promising tools for predicting respiratory hospitalizations but still require greater methodological standardization, integration of multiple data sources, and development of more explainable models adaptable to different urban realities.

Keywords: Respiratory diseases; Air pollution; Machine learning; Predictive modeling; Urban public health.

Sumário

1 Introdução **1**

1.1 Contextualização 1

1.2 Motivação 2

1.3 Objetivos 2

1.4 Metodologia 3

1.5 Organização dos Capítulos 3

2 Fundamentação teórica **5**

2.1 Dados abertos do SUS 5

2.2 Classificação Internacional de Doenças 6

2.3 Air Quality Index 7

2.4 Séries Temporais 7

2.4.1 Definição e tipos de séries 7

2.4.2 Frequência e granularidade temporal 8

2.4.3 Componentes das séries temporais 9

2.4.4 Estacionariedade, dependência temporal e autocorrelação 9

2.4.5 Horizonte de previsão e *lead time* 11

2.5 Modelo Supervisionado de Aprendizado de Máquina 12

2.5.1 Random Forest 13

2.6 Interpretabilidade de modelos e SHAP 14

3 Trabalhos Relacionados **16**

3.1 Estratégia de Seleção e Priorização de Trabalhos Relacionados 16

3.1.1 Critérios de inclusão 17

3.1.2 Critérios de Exclusão 18

3.1.3 Critérios de Priorização 18

3.1.4 Resultados 18

3.2 Comparação entre os Trabalhos Relacionados 19

3.2.1 Modelos de Aprendizado de Máquina Utilizados 19

3.2.2 Variáveis Ambientais e Meteorológicas Mais Relevantes 20

3.2.3 Desfechos de Saúde Considerados 21

3.2.4 Estratégias de Explicabilidade, Otimização e Tratamento dos Dados 21
3.3 Discussão 22
3.3.1 Síntese crítica das convergências 22
3.3.2 Divergências e diferenças metodológicas 22
3.3.3 Lacunas identificadas 23
3.3.4 Implicações para o trabalho proposto 23
3.3.5 Considerações sobre o contexto brasileiro 24
3.3.6 Síntese final 24

4 Metodologia de Integração de Dados, Engenharia de Atributos e Modelagem Preditiva 26
4.1 Formulação do Problema e Justificativa do Modelo 26
4.2 Visão Geral do Fluxo Metodológico 27
4.3 Obtenção e Integração das Bases de Dados 29
4.3.1 Dados de Internações Hospitalares 29
4.3.2 Dados de Qualidade do Ar e Meteorologia 30
4.3.3 Critérios de Seleção, Filtragem e Recorte Temporal. 32
4.3.4 Integração das Fontes e Construção da Base Diária Unificada 32
4.4 Pré-processamento das Séries Temporais 33
4.4.1 Tratamento de Ausências e Consistência Temporal 34
4.4.2 Agregação para Frequência Diária 36
4.4.3 Transformações Estatísticas (Yeo–Johnson) 37
4.4.4 Unificação das Bases e Construção do *Dataset* Final 40
4.5 Análises Temporais Exploratórias 41
4.5.1 Análise da Série de Internações 41
4.5.2 Dependência Temporal do Alvo 45
4.5.3 Análise das Variáveis Ambientais 47
4.6 Análise das Relações entre Variáveis Ambientais e Internações 49
4.6.1 Correlação Instantânea (Lag 0) 49
4.6.2 Correlações Defasadas (*lags* ambientais) 50
4.6.3 Médias Móveis Defasadas e Exposição Acumulada 51
4.6.4 Interpretação dos Padrões Observados 52
4.7 Engenharia de Atributos 53

4.7.1 Features Autorregressivas das Internações (*lags* + médias móveis) 53
4.7.2 *Features* de Sazonalidade (mês, semana, codificação cíclica) 54
4.7.3 *Features* da Decomposição STL (componente anual e amplitude) 55
4.7.4 *Features* Meteorológicas e de Poluentes 55
4.7.5 Construção Final da Matriz de Features 56
4.8 Desenvolvimento do Modelo Preditivo 57
4.8.1 Estratégia de Treino/Teste e Validação Temporal 57
4.8.2 *Detrending* Causal do Alvo 58
4.8.3 Seleção de Variáveis 58
4.8.4 Treinamento, Ajuste de Hiperparâmetros e Previsão 59
4.8.5 Métricas de Avaliação e Procedimentos Diagnósticos 59

5 Resultados 60
5.1 Plano de Experimentos 60
5.1.1 Objetivo dos experimentos 60
5.1.2 Descrição dos cenários testados 61
5.1.3 Conjunto de treino, teste e configuração temporal 61
5.1.4 Análises complementares 62
5.2 Resultados do modelo sem `detrend` 62
5.3 Resultados do modelo com `detrend` 64

6 Conclusões 67
6.1 Análise Retrospectiva 67
6.2 Trabalhos Futuros 68

Referências 69

Lista de Figuras

FIGURA 1: Exemplo de série univariada e multivariada 8

FIGURA 2: Série temporal simulada com tendência e sazonalidade 10

FIGURA 3: Componente de tendência da série temporal 10

FIGURA 4: Componente de sazonalidade da série temporal 11

FIGURA 5: Componente de ruído (resíduo) da série temporal 11

FIGURA 6: Componente de ciclo de uma série temporal 12

FIGURA 7: Esquema do modelo Random Forest 14

FIGURA 8: Fluxograma feito com PRISMA para seleção dos artigos. 19

FIGURA 9: Fluxo metodológico do estudo 29

FIGURA 10: Distribuição do NO antes da transformação 38

FIGURA 11: Q-Q Plot do NO 38

FIGURA 12: Distribuição do NO após Yeo-Johnson 39

FIGURA 13: Q-Q Plot do NO após Yeo-Johnson 40

FIGURA 14: Série temporal das internações 42

FIGURA 15: Perfil mensal das internações 43

FIGURA 16: Perfil semanal das internações 44

FIGURA 17: Decomposição *Seasonal and Trend decomposition using Loess* (STL) da série de internações 45

FIGURA 18: Função de Autocorrelação Parcial (PACF) das internações 46

FIGURA 19: Correlação entre internações e médias móveis 47

FIGURA 20: *Boxplot* das variáveis meteorológicas e poluentes 48

FIGURA 21: *Boxplots* após transformação das variáveis 48

FIGURA 22: Correlação Spearman (*lag* 0) entre ambiente e internações. 50

FIGURA 23: Correlação defasada entre poluentes e internações 51

FIGURA 24: Correlação entre internações e médias móveis defasadas 52

FIGURA 25: Real vs previsto no teste (sem `detrend`) 63

FIGURA 26: Resumo *Shapley Additive Explanations* (SHAP) (sem `detrend`) 64

FIGURA 27: Real vs previsto (*Random Forest* (RF) com *detrend*) 65

FIGURA 28: SHAP summary (RF com *detrend*) 66

Lista de Tabelas

TABELA 1: Capítulo X da Classificação Internacional de Doenças (CID) — Doenças respiratórias 6

TABELA 2: Faixas e categorias do Índice de Qualidade do Ar (IQAR) 7

TABELA 3: Comparação entre estudos sobre predição de doenças respiratórias 25

TABELA 4: Estações e poluentes monitorados 31

TABELA 5: Formato final dos dados de internações. 33

TABELA 6: Série diária de internações 36

TABELA 7: Formato final dos dados meteorológicos e atmosféricos 37

TABELA 8: Estrutura do *dataset* diário (variáveis ambientais) 40

TABELA 9: Métricas do RF sem `detrend` 62

TABELA 10: Desempenho do RF com `detrend` 65

LISTA DE ABREVIAÇÕES

ACF	Função De Autocorrelação	9, 10
ADF	*Augmented Dickey–Fuller*	10
AIH	Autorização De Internação Hospitalar	29, 32
ANS	Agência Nacional De Saúde Suplementar	5
AQI	Air Quality Index	7, 31, 32
CID	Classificação Internacional De Doenças	xii, 6, 32
CID-10	CID 10^{A} Revisão	5
CSV	*Comma-Separated Values*	29, 34, 37
DATARIO	Portal DATA.RIO Da Prefeitura Do Rio De Janeiro	7, 27, 29, 31, 67
DATASUS	Departamento De Informática Do SUS	5, 6, 29
ELM	*Extreme Learning Machines*	19, 25
EPA	*United States Environmental Protection Agency*	7, 32
GLM	*Generalized Linear Model* (Modelo Linear Generalizado)	20
IQAR	Índice De Qualidade Do Ar	7
KPSS	Kwiatkowski–Phillips–Schmidt–Shin	10
LIME	*Local Interpretation through Model-agnostic Explanations* (Lógica Interpretável Para Modelos Explicáveis)	21, 23
LST	*Land Surface Temperature*	20, 25
LSTM	*Long Short-Term Memory*	20, 25
MAE	*Mean Absolute Error* (Erro Absoluto Médio)	3, 23, 59, 62, 64, 65
MLP	*Multilayer Perceptron*	19, 25
NDVI	*Normalized Difference Vegetation Index*	20, 25
OMS	Organização Mundial Da Saúde	32
PACF	Função De Autocorrelação Parcial	xi, 9, 10, 45, 46, 63
PDP	*Partial Dependence Plot*	23
PFI	*Permutation Feature Importance*	23
PSO	*Particle Swarm Optimization*	21
PYSUS	*Python SUS* (Biblioteca Para Acesso A Dados Do DATASUS)	6, 29, 67
R2	Coeficiente De Determinação	3
RBF	*Radial Basis Function*	20, 25

RD	Registro De Internação Hospitalar	29, 32
RF	*Random Forest*	xi, 3, 10, 13, 14, 15, 16, 19, 23, 25, 27, 28, 58, 59, 60, 62, 63, 64, 65, 66, 67, 68
RMSE	*Root Mean Squared Error* (Raiz Do Erro Quadrático Médio)	3, 23, 59, 62, 64, 65
SHAP	*Shapley Additive Explanations*	xi, 3, 5, 14, 15, 21, 23, 24, 25, 60, 62, 63, 64, 65, 66, 67
SIH	Sistema De Informações Hospitalares	2, 26, 27, 28, 29, 33, 36
SIH/SUS	Sistema De Informações Hospitalares Do SUS	5, 6, 67
SMAPE	*Symmetric Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio Simétrico)	3, 23, 59, 62, 65
STL	*Seasonal and Trend decomposition using Loess*	xi, 3, 23, 26, 44, 45, 53, 55, 56, 57, 61, 66, 67, 68
SUS	Sistema Único De Saúde	5, 6, 24, 26, 27, 28, 29, 33, 36
SVR	*Support Vector Regression*	20, 25
TCC	Trabalho De Conclusão De Curso	3, 16, 18, 23
WMAPE	*Weighted Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio Ponderado)	3, 23, 59, 62, 65
XAI	*Explainable artificial intelligence*	23, 24

Declaração de Uso de Ferramentas de Inteligência Artificial

O autor declara que utilizou ferramentas de Inteligência Artificial generativa exclusivamente como apoio à produção textual desta dissertação, incluindo sugestões de redação, revisão linguística, esclarecimento de conceitos técnicos e organização de ideias. Todas as decisões conceituais, análises, interpretações, escolhas metodológicas, resultados experimentais e conclusões apresentadas são de responsabilidade integral do autor. Nenhum conteúdo gerado por ferramentas de Inteligência Artificial foi incorporado sem revisão crítica. O uso dessas ferramentas ocorreu em conformidade com princípios de integridade acadêmica, transparência e responsabilidade, e não substituiu o trabalho intelectual necessário à elaboração desta pesquisa.

Capítulo 1

Introdução

1.1 Contextualização

A poluição atmosférica é reconhecida como um dos principais fatores de risco ambientais para a saúde pública, especialmente em grandes centros urbanos de países em desenvolvimento. Segundo o *World Health Statistics 2025*, aproximadamente 6,7 milhões de mortes em 2019 foram atribuídas à exposição a material particulado fino [World Health Organization, 2025].

Os principais poluentes atmosféricos que afetam a saúde humana incluem material particulado fino ($PM_{2.5}$ e PM_{10}), dióxido de enxofre (SO_2), dióxido de nitrogênio (NO_2), monóxido de carbono (CO) e ozônio (O_3). Esses compostos estão associados a diversos efeitos adversos, especialmente entre grupos vulneráveis como crianças, idosos e pessoas com comorbidades [Ministério da Saúde, 2021].

Estudos internacionais reforçam a relação entre poluição atmosférica e desfechos respiratórios. Dowlatabadi et al. [2024], por exemplo, encontraram em Mashhad, no Irã, uma forte correlação entre concentrações elevadas de PM_{10} e NO_2 e o aumento da mortalidade por doenças respiratórias. Semelhantemente, Rautela and Goyal [2025] observaram no subcontinente indiano associações significativas entre a exposição ao $PM_{2.5}$ e a mortalidade por doenças respiratórias crônicas. Embora muitos desses trabalhos se concentrem em mortalidade, as mesmas relações de risco são relevantes para internações, foco deste estudo.

No Brasil, estima-se que cerca de 326.478 óbitos entre 2019 e 2021 estejam associados à exposição a poluentes atmosféricos ao ar livre, atingindo principalmente a população idosa [Buralli and Connerton, 2025]. Estudos nacionais, como os de Miranda et al. [2021], Kachba et al. [2020], Araujo et al. [2020] e Reis et al. [2025], também investigaram a relação entre qualidade do ar e internações respiratórias, explorando diferentes técnicas analíticas. Embora heterogêneos em recortes temporais, modelos e regiões estudadas, esses trabalhos evidenciam o potencial de integrar dados ambientais e de saúde.

Apesar dos avanços, ainda há escassez de estudos dedicados à realidade urbana do município do Rio de Janeiro, cujas particularidades climáticas, urbanísticas e de cobertura de monitora-

mento demandam análises próprias.

1.2 Motivação

Embora o Brasil possua sistemas públicos robustos de coleta de dados ambientais e hospitalares, como o MonitorAr e o Sistema de Informações Hospitalares (SIH), o uso integrado dessas bases com fins preditivos ainda é pouco explorado. A literatura internacional demonstra avanços significativos na previsão de internações respiratórias com aprendizado de máquina, mas esses estudos concentram-se majoritariamente em países com infraestrutura de monitoramento mais homogênea e dados clínicos padronizados. No contexto brasileiro, persistem lacunas importantes: ausência de protocolos que conciliem múltiplos poluentes e variáveis meteorológicas em frequência diária, falta de documentação explícita sobre tratamento de ausências e sazonalidade, e escassez de análises que incorporem explicabilidade operacional para uso em saúde pública.

A cidade do Rio de Janeiro reúne condições que tornam essa análise especialmente relevante: elevada densidade populacional, diversidade climática entre regiões, variações locais na emissão de poluentes e forte dependência do sistema público de saúde. Entretanto, enfrenta também limitações estruturais, como estações que não monitoram todos os poluentes e lacunas de medição que dificultam análises diretas.

Este trabalho busca preencher essas lacunas ao propor uma abordagem integrada que conecta dados ambientais e hospitalares em escala diária, trata defasagens e sazonalidade de forma transparente e incorpora mecanismos de explicabilidade para uso em saúde pública. Ao endereçar esses desafios, o estudo contribui para aproximar a pesquisa aplicada da realidade operacional do SUS, oferecendo evidências para planejamento hospitalar e formulação de políticas públicas em contextos urbanos complexos.

1.3 Objetivos

Este trabalho propõe uma abordagem preditiva baseada em aprendizado de máquina supervisionado para antecipar o número diário de hospitalizações por doenças respiratórias na cidade do Rio de Janeiro, a partir da integração de dados hospitalares e ambientais.

Os objetivos específicos incluem:

1. Integrar dados hospitalares (SIH/SUS) e ambientais (MonitorAr/DATA.RIO) em uma

base diária unificada;

2. Aplicar estratégias de pré-processamento com tratamento de ausências, agregação temporal e transformação estatística (Yeo–Johnson);

3. Criar atributos autorregressivos, sazonais (incluindo codificação cíclica) e derivados da decomposição STL (Seasonal and Trend Decomposition Using Loess);

4. Incorporar variáveis ambientais contemporâneas, defasadas e acumuladas (médias móveis causais);

5. Desenvolver e avaliar modelos preditivos com *Random Forest*, incluindo ajuste de hiperparâmetros, *detrending* causal do alvo e análise de interpretabilidade via SHAP.

1.4 Metodologia

A metodologia é estruturada em quatro etapas principais: (i) integração das bases; (ii) pré-processamento com imputação restrita e transformação Yeo–Johnson; (iii) engenharia de atributos (*lags*, médias móveis, codificação cíclica, STL e variáveis ambientais defasadas/acumuladas); e (iv) modelagem preditiva com RF, validação temporal e métricas expandidas (*Mean Absolute Error* (Erro Absoluto Médio) (MAE), *Root Mean Squared Error* (Raiz do Erro Quadrático Médio) (RMSE), Coeficiente de determinação (R2), *symmetric Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio Simétrico) (sMAPE), *Weighted Mean Absolute Percentage Error* (Erro Percentual Absoluto Médio Ponderado) (wMAPE) e *Bias*). Além disso, adota-se um horizonte de previsão de 1 dia a frente, garantindo que todas as variáveis utilizadas sejam estritamente causais. Essa escolha se deve ao objetivo de avaliar relações de curto prazo entre exposição ambiental recente e internações subsequentes, evitando qualquer vazamento de informação do dia previsto.

1.5 Organização dos Capítulos

Este trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica sobre os conceitos essenciais para a pesquisa. O Capítulo 3 aborda os artigos relacionados ao tema, mostrando uma avaliação daqueles que melhor contribuem cientificamente para o objetivo do Trabalho de Conclusão de Curso (TCC). No Capítulo 4, são detalhados os procedimentos metodológicos, incluindo as etapas de coleta, processamento dos dados e construção

dos modelos. O Capítulo 5 apresenta os resultados das análises realizadas. O Capítulo 6 discute as conclusões e perspectivas futuras.

Capítulo 2

Fundamentação teórica

Neste capítulo, apresentamos os principais conceitos utilizados para o desenvolvimento da metodologia deste trabalho. Na Seção 2.1, explicamos os dados abertos de saúde disponibilizados pelo Sistema Único de Saúde (SUS). Na Seção 2.2, abordamos a classificação das doenças respiratórias segundo a Classificação Internacional de Doenças, especificamente a CID 10ª Revisão (CID-10). Na Seção 2.3, explicamos o índice para qualidade do ar. Na seção 2.4, apresentamos os conceitos relacionados a séries temporais. Na Seção 2.5, detalhamos sobre o modelo supervisionado de aprendizado de máquina utilizado para construir uma solução para o problema proposto pelo presente trabalho. Por fim, na seção 2.6, explicamos o método SHAP, empregado para quantificar a contribuição de cada variável nas previsões do modelo.

2.1 Dados abertos do SUS

O SUS disponibiliza, por meio do Departamento de Informática do SUS (DATASUS), um amplo conjunto de dados públicos com atualizações regulares relacionados à saúde da população brasileira. Esses dados podem subsidiar análises objetivas da situação sanitária do país, apoiar a elaboração de programas e políticas públicas, e fomentar a tomada de decisão baseada em evidências[1]. Assim, o conjunto de dados organiza informações de saúde por áreas temáticas, entre elas: estatísticas vitais, epidemiologia, morbidade, assistência à saúde, recursos financeiros, dados demográficos e socioeconômicos, além de links para informações relacionadas à saúde suplementar, via Agência Nacional de Saúde Suplementar (ANS).

No contexto desta pesquisa, são utilizados em específico dados de morbidade hospitalar, disponibilizados pelo Sistema de Informações Hospitalares do SUS (SIH/SUS), acessível na seção "Epidemiológicas e Morbidade" do `Tabnet`. Essa base contempla registros administrativos de internações hospitalares realizadas no âmbito do SUS em todo o território nacional. Os dados podem ser consultados e agregados por diferentes critérios, como local de internação ou local de residência do paciente, além de contemplarem internações por causas gerais ou por causas

[1] Disponível em: `http://www2.datasus.gov.br`

externas (como acidentes e violências). As séries históricas disponíveis variam conforme a segmentação escolhida, podendo abranger registros de 2008 até os dias atuais, ou, em alguns casos, de períodos anteriores, como de 1984 a 2007 ou dos anos 1990 a 2007. Além disso, os dados são estruturados em arquivos padronizados com o formato `.dbc`, o qual é próprio do sistema.

Embora o acesso aos dados do SIH/SUS possa ser feito diretamente pela interface web do`Tabnet`, também é possível automatizar a obtenção e o tratamento das bases utilizando bibliotecas específicas em *Python*. Para isso, existe a biblioteca *Python SUS* (biblioteca para acesso a dados do DATASUS) (PySUS) [2], que permite o download, leitura e manipulação de arquivos do DATASUS via código em um ambiente Linux ou pelo *Google Colab*, simplificando, assim, o processo de extração e análise dos dados de morbidade.

2.2 Classificação Internacional de Doenças

A CID, mantida pela OMS, fornece uma classificação para diferentes doenças que atingem o ser humano, com termos clínicos que podem ser utilizados para auxiliar na manutenção de registros de saúde [World Health Organization, 2019]. O uso da CID é essencial para se realizar uma filtragem dos dados do SUS no, para retornar somente doenças de natureza respiratória. Para tanto, utilizamos o capítulo X da Classificação, o qual se divide em dez grupos, descritos na Tabela 1.

Tabela 1: Classificação do Capítulo X da CID, que agrupa os códigos relacionados às Doenças do Aparelho Respiratório. A subdivisão contempla infecções agudas, doenças crônicas, doenças da pleura, entre outras categorias que afetam os pulmões e vias aéreas.

Código	Grupo de doenças
J00–J06	Infecções agudas das vias aéreas superiores
J09–J18	Influenza [gripe] e pneumonia
J20–J22	Outras infecções agudas das vias aéreas inferiores
J30–J39	Outras doenças das vias aéreas superiores
J40–J47	Doenças crônicas das vias aéreas inferiores
J60–J70	Doenças pulmonares devidas a agentes externos
J80–J84	Outras doenças respiratórias que afetam principalmente o interstício pulmonar
J85–J86	Afecções necróticas e supurativas das vias aéreas inferiores
J90–J94	Outras doenças da pleura
J95–J99	Outras doenças do aparelho respiratório

Fonte: Ministério da Saúde (2008), disponível em `http://www2.datasus.gov.br/cid10/V2008/cid10.htm`.

[2] Disponível em: `https://github.com/AlertaDengue/PySUS`

2.3 Air Quality Index

O Air Quality Index (AQI) foi desenvolvido pela *United States Environmental Protection Agency* (EPA) visando comunicar a qualidade do ar em ambientes ao ar livre para a população de uma maneira intuitiva [United States Environmental Protection Agency, nd]. A versão brasileira do índice, denominada Índice de Qualidade do Ar (IQAR), mantido no município do Rio de Janeiro pelo DATA RIO, apresenta um valor numérico capaz de identificar a qualidade do ar com base nas concentrações de poluentes registrados em dado período [Prefeitura do Rio de Janeiro, 2024a], incluindo partículas finas ($PM_{2.5}$, PM_{10}), monóxido de carbono, dióxido de enxofre, dióxido de nitrogênio e ozônio. Quanto maior for o valor, pior a qualidade do ar e o risco à saúde da população. A Tabela 2 mostra o índice AQI e o nível de preocupação relacionado, assim como a descrição da qualidade do ar nesse intervalo, conforme a versão adaptada do Portal DATA.RIO da Prefeitura do Rio de Janeiro (DATARIO).

Tabela 2: Categorias utilizadas no IQAR, suas faixas numéricas e respectivas descrições.

Categoria	Faixa IQAR	Descrição da qualidade do ar
Boa	0–50	Satisfatória e apresenta pouco ou nenhum risco para a saúde da população.
Regular	51–100	Aceitável, mas pode causar preocupação moderada à saúde de indivíduos extremamente sensíveis.
Inadequada	101–199	Indivíduos de grupos sensíveis podem apresentar efeitos à saúde. A população em geral, no entanto, não costuma ser afetada.
Má	200–299	Toda a população começa a sentir efeitos à saúde.
Péssima	300 ou mais	Situação crítica, com potencial de efeitos adversos à saúde para toda a população; recomenda-se alerta.

Fonte: Adaptado de informações oficiais sobre qualidade do ar. Acesso em: 25 jun. 2025.

2.4 Séries Temporais

2.4.1 Definição e tipos de séries

Séries temporais são conjuntos de observações de dados coletados ao longo de intervalos regulares de tempo, ordenados de forma cronológica. Em geral, esses conjuntos são igualmente espaçados no tempo [Haben et al., 2023]. Isso possibilita realizar análises dos padrões nos

dados observados para extrair informações relevantes como tendências, ciclos e sazonalidade, e prever valores futuros com base nos padrões históricos.

As séries temporais podem ser classificadas como univariadas ou multivariadas, dependendo do número de variáveis observadas ao longo do tempo. As séries univariadas registram apenas uma variável ao longo do tempo, como a temperatura diária de uma cidade ou o número de internações hospitalares por dia. Já as multivariadas, por sua vez, possuem múltiplas variáveis medidas simultaneamente em cada ponto do tempo. Um exemplo visual de como esses dois tipos de séries se comportam pode ser visto na Figura 1. Esse comportamento das séries multivariadas é especialmente interessante porque consegue mostrar possíveis dependências entre as variáveis medidas [Nielsen, 2019]. Por exemplo, quando tentamos prever internações hospitalares, podemos incluir variáveis como temperatura, umidade e poluentes para identificar padrões e verificar como elas se relacionam.

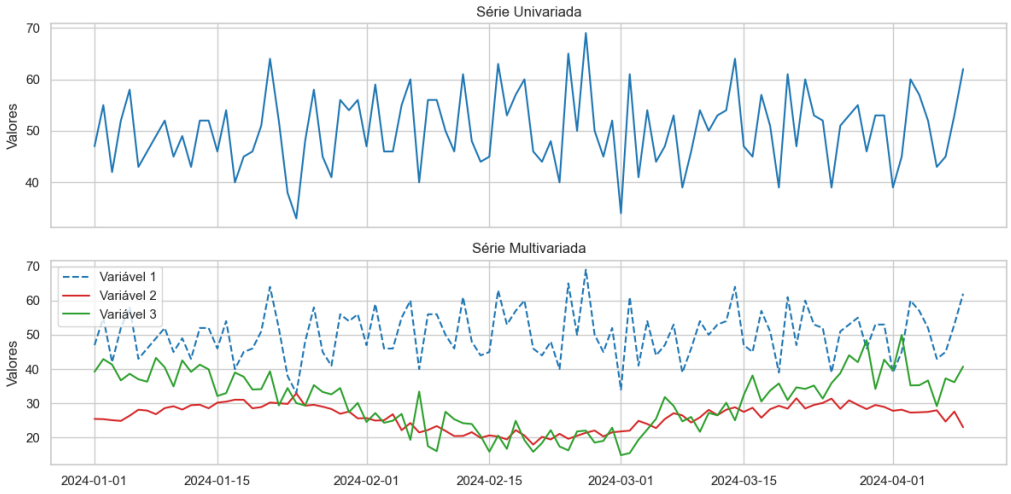

Figura 1: Exemplo comparativo entre uma série univariada e uma série multivariada ao longo de um certo período de tempo.

2.4.2 Frequência e granularidade temporal

Outro aspecto importante a ser considerado em uma série temporal é a sua frequência temporal, que diz respeito à periodicidade com que as observações são registradas, podendo ser horária, diária, semanal, mensal, entre outras. Essa frequência influencia diretamente a granularidade da análise e a escolha do modelo preditivo, já que padrões de sazonalidade, tendência e ruído variam conforme o intervalo entre os dados. Assim, séries de alta frequência capturam variações rápidas e ruído, enquanto séries de baixa frequência tendem a enfatizar ciclos

sazonais e tendências de longo prazo. Uma série mensal de internações hospitalares pode, por exemplo, captar padrões diferentes de uma série diária, o que pode exigir abordagens distintas para a modelagem.

2.4.3 Componentes das séries temporais

Uma série temporal pode ser decomposta em diferentes componentes que ajudam a explicar o seu comportamento ao longo do tempo. Os principais são: tendência, sazonalidade, ciclos e ruído. Essa decomposição é útil tanto para realizar análises quanto para realizar uma modelagem preditiva por permitir isolar padrões estruturais dos efeitos aleatórios.

Tendência refere-se ao comportamento de longo prazo da série, indicando se há um aumento, queda ou estabilidade nos valores ao longo do tempo, podendo ser visualizado pela curva suavizada ao longo do tempo, como médias móveis ou regressão polinomial. Sazonalidade, por sua vez, representa padrões que se repetem em intervalos regulares e previsíveis, como variações mensais, trimestrais ou anuais. Um ciclo refere-se a flutuações de médio a longo prazo que não têm periodicidade fixa, mas ainda assim mostram um comportamento repetitivo. Diferentemente da sazonalidade, os ciclos estão geralmente associados a fatores econômicos, sociais ou estruturais mais amplos. Além disso, considera-se também o ruído, que representa a variação imprevisível dos dados, causada por fatores aleatórios ou não mensuráveis. Esse componente é geralmente interpretado como o resíduo da série, e idealmente deve ser minimizado pelos modelos preditivos.

Muitas séries podem incluir um ou mais desses componentes. Dessa forma, para escolher um método de previsão, é preciso identificar primeiro os padrões nos dados e em seguida escolher o método apropriado [Hyndman and Athanasopoulos, 2018]. Para ilustrar o comportamento desses componentes, na Figura 2, temos um exemplo de uma série temporal simulada. A decomposição dessa série que mostra a sua tendência, sazonalidade e ruídos pode ser vista respectivamente na Figura 3, Figura 4 e Figura 5. Além disso, para ilustrar a presença de um ciclo, temos a Figura 6

2.4.4 Estacionariedade, dependência temporal e autocorrelação

Quando não há tendência ou sazonalidade que altere a estrutura da série conforme o tempo avança, dizemos que a série é estacionária. A Função de autocorrelação (ACF) e PACF é útil

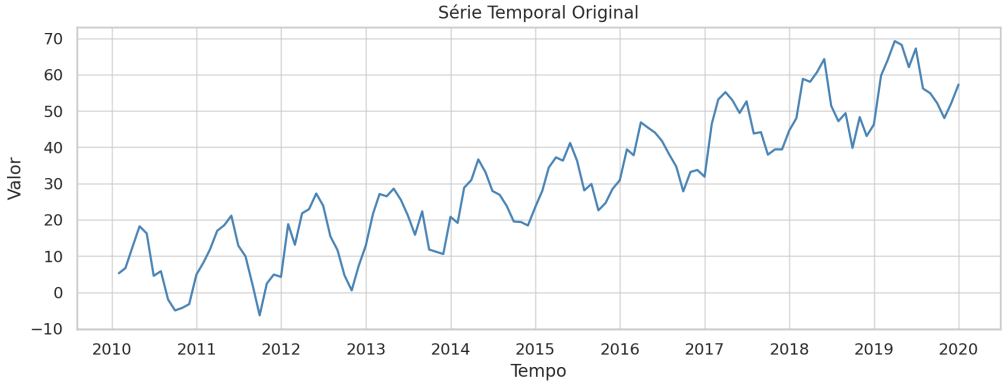

Figura 2: Série temporal sintética simulada com 120 observações mensais. A série foi composta pela combinação de três elementos: tendência linear crescente, padrão sazonal com periodicidade anual e ruído aleatório.

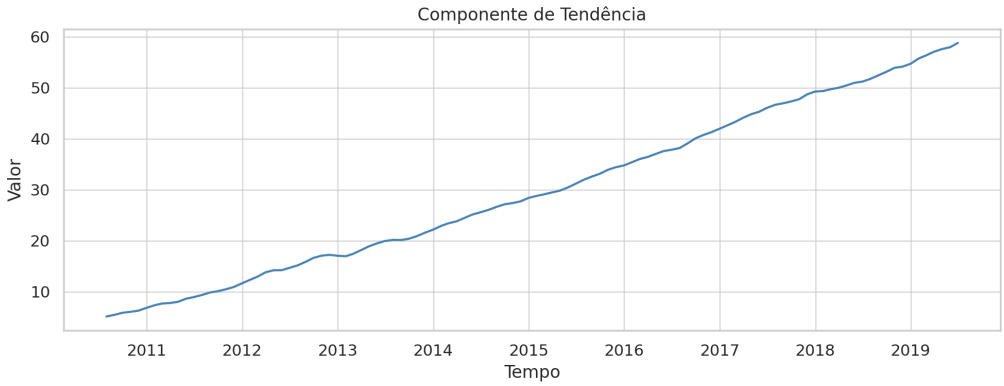

Figura 3: Componente de tendência extraído a partir da decomposição aditiva da série temporal. Indica o comportamento de longo prazo da série, mostrando um crescimento consistente ao longo do tempo.

para identificar dependências temporais e selecionar defasagens relevantes [Leites et al., 2025; Widodo et al., 2016]. A ACF mede a correlação entre a série e suas próprias defasagens ao longo do tempo, enquanto a PACF elimina os efeitos intermediários e identifica as defasagens com influência direta. Embora testes formais de estacionariedade, como *Augmented Dickey–Fuller* (ADF) ou Kwiatkowski–Phillips–Schmidt–Shin (KPSS), sejam comuns em abordagens clássicas [Dar et al., 2024], modelos baseados em árvores, como RF, não exigem estacionarização explícita. Em vez disso, neste trabalho, a seleção de defasagens relevantes foi baseada apenas na análise do PACF.

Essas defasagens são chamadas de variáveis defasadas (*lags*) e podem ser utilizadas como preditores em modelos supervisionados. Então, por exemplo, para prever um valor y_t, pode-se usar como entrada os valores $y_{t-1}, y_{t-2}, \ldots, y_{t-k}$, onde k é o número de defasagens escolhidas com base na análise da função de autocorrelação usada.

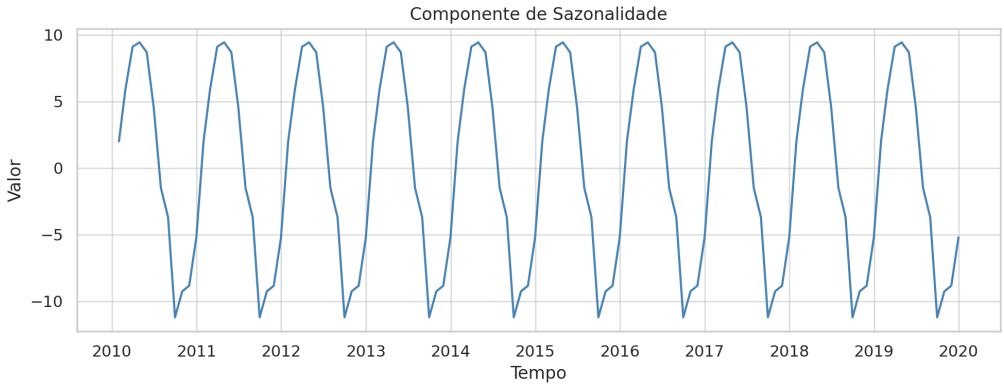

Figura 4: Componente de sazonalidade resultante da decomposição da série. Reflete padrões periódicos que se repetem a cada 12 meses, caracterizando a influência sazonal nos dados.

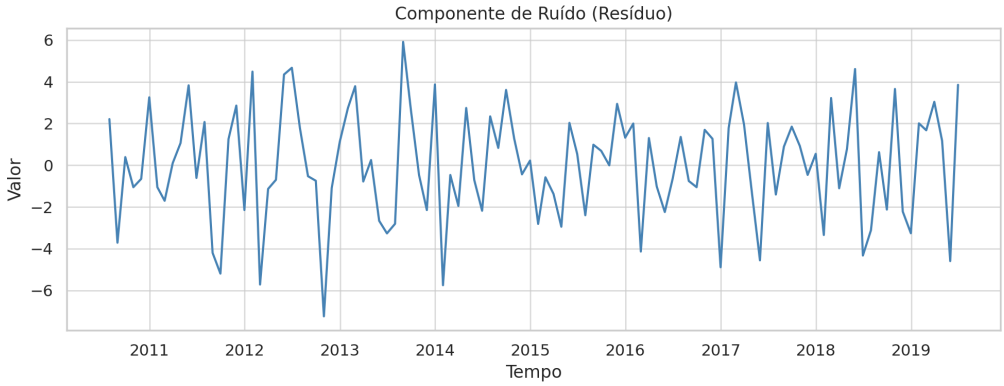

Figura 5: Componente de ruído ou resíduo, que representa as variações imprevisíveis da série após a remoção da tendência e da sazonalidade. Normalmente associadas a flutuações aleatórias e erros de medição.

Além disso, para estruturar os dados com essas defasagens, utiliza-se a técnica de janela deslizante, que percorre a série com uma janela fixa de tamanho k. Em cada passo, essa janela extrai os k valores anteriores como entrada e associa ao valor seguinte como saída. Assim, o uso dessa estratégia transforma a série temporal em um conjunto de pares entrada–saída, compatível com modelos supervisionados como regressão, árvores ou redes neurais. Neste trabalho, entretanto, não empregamos a técnica de janela deslizante no sentido clássico, optando por construir apenas defasagens pontuais e médias móveis, que também são uma prática consolidada para melhorar a capacidade preditiva, conforme sugerido por Joseph et al. [2024].

2.4.5 Horizonte de previsão e *lead time*

O horizonte de previsão define quantos passos à frente se deseja estimar, sendo um conceito central em tarefas de previsão temporal [Haben et al., 2023]. Por exemplo, em uma série diária,

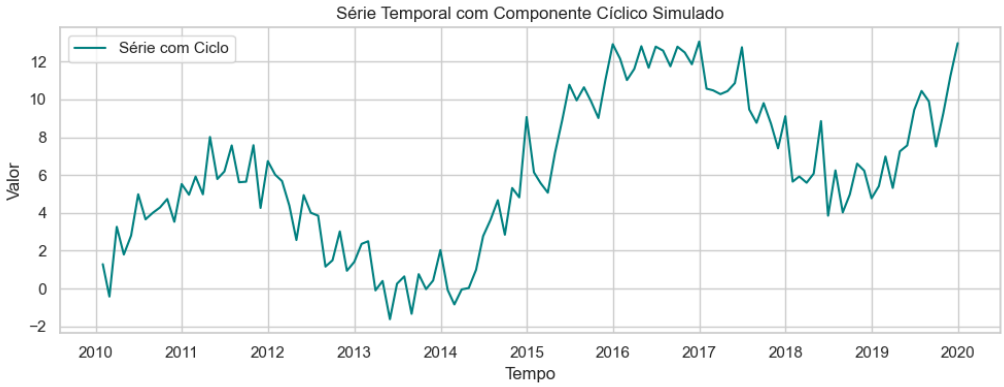

Figura 6: Componente de ciclo de uma série temporal, representando flutuações periódicas de longo prazo não relacionadas à sazonalidade. Observe que o ciclo difere da sazonalidade por sua irregularidade e períodos não fixos.

um horizonte de previsão igual a 1 indica a previsão do próximo dia; já um horizonte de 7 se refere à estimativa para uma semana adiante. Assim, quanto maior for o horizonte, maior tende a ser a incerteza associada, especialmente em séries com alto grau de variabilidade ou eventos sazonais. Por isso, a escolha do horizonte de previsão impacta diretamente o desenho da tarefa supervisionada e a dimensão da janela deslizante usada como entrada. Em problemas com horizonte curto, é possível capturar dependências diretas e recentes da série; em horizontes longos, a influência de observações mais antigas pode se tornar mais relevante ou mais difícil de modelar. Para a metodologia, portanto, escolhemos trabalhar apenas com um horizonte de previsão de 1 dia.

Outro conceito relacionado é o *lead time,* que representa o intervalo entre o momento em que a previsão é feita e o instante em que a informação prevista será efetivamente utilizada. Esse intervalo pode surgir por restrições operacionais, logísticas ou de coleta de dados. Por exemplo, em um sistema de saúde, pode ser necessário prever o número de internações com dois dias de antecedência para alocar recursos, exigindo um modelo capaz de fornecer estimativas com esse *lead time*.

2.5 Modelo Supervisionado de Aprendizado de Máquina

O aprendizado de máquina supervisionado é uma abordagem na qual o modelo é treinado a partir de um conjunto de dados rotulado, ou seja, com entradas e saídas conhecidas, respectivamente denominadas como variáveis preditoras e variáveis preditivas. O objetivo é aprender uma função que relacione essas entradas e saídas para posteriormente realizar previsões sobre

novos dados.

Entre os algoritmos supervisionados mais utilizados em tarefas de regressão e classificação, selecionamos o RF para abordar na tarefa de previsão de hospitalizações, descrito a seguir.

2.5.1 Random Forest

Árvores de decisão são modelos clássicos de aprendizado supervisionado não paramétrico, capazes de capturar relações não lineares entre variáveis e gerar regras interpretáveis de classificação ou regressão [Ji et al., 2023]. No entanto, quando utilizadas isoladamente, essas árvores podem apresentar alta variância e sensibilidade ao conjunto de treinamento. Para então, superar essa limitação, foi proposto o modelo RF, o qual combina diversas árvores de decisão para aumentar a robustez e reduzir o risco de *overfitting*.

O RF é um algoritmo de *ensemble learning*, ou seja, um método que agrupa múltiplos modelos fracos para formar um modelo mais forte. Assim, a ideia central é treinar várias árvores de decisão, cada uma em uma amostra diferente do conjunto de dados, obtida por meio da técnica de *bootstrapping*, a qual é uma amostragem aleatória com reposição. Isso garante diversidade entre as árvores, uma vez que cada uma é treinada com subconjuntos distintos dos dados originais.

Além disso, no momento de construir cada árvore, o algoritmo seleciona aleatoriamente apenas um subconjunto de variáveis disponíveis para considerar em cada divisão (*split*). Essa aleatoriedade adicional evita que todas as árvores tomem decisões muito parecidas, promovendo diversidade estrutural entre elas.

Por fim, durante a etapa de inferência, as previsões das árvores são combinadas por meio de uma regra de agregação. Em tarefas de classificação, adota-se o voto da maioria; já em regressão, utiliza-se a média das previsões de todas as árvores. Essa combinação reduz a variância do modelo sem aumentar significativamente o viés, resultando em um preditor mais estável e preciso.

Como ilustra a Figura 7, o modelo RF se baseia na diversidade entre árvores individuais para construir uma previsão agregada mais robusta. Essa abordagem tem se mostrado eficaz em diversos estudos que exploram a previsão de internações hospitalares ou doenças respiratórias a partir de dados ambientais e meteorológicos [Ravindra et al., 2023; Cappelli et al., 2024; Barnett-Itzhaki et al., 2025; Yang et al., 2025].

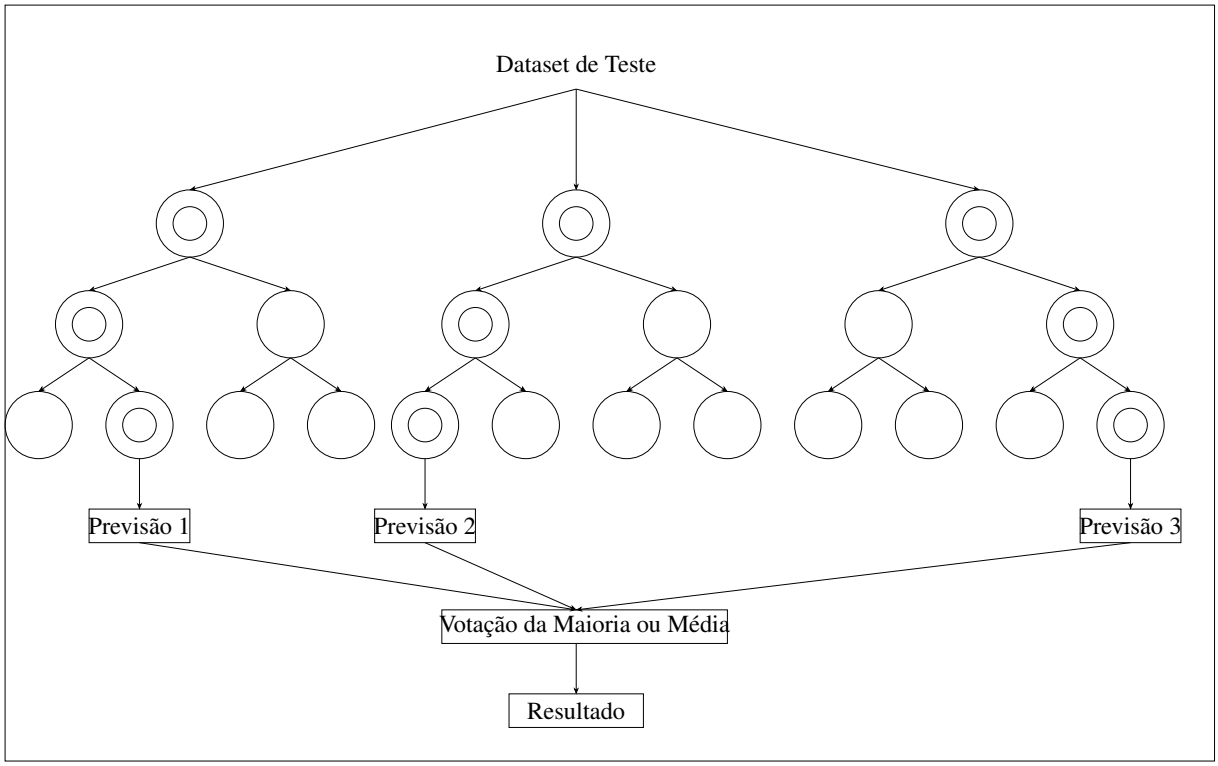

Figura 7: Representação esquemática do funcionamento do modelo *Random Forest*. O algoritmo constrói diversas árvores de decisão independentes a partir de subconjuntos aleatórios dos dados de treino e das variáveis preditoras (método de *bagging*). A predição final é obtida por agregação, como a média, no caso de regressão, ou o voto majoritário, no caso de classificação.

2.6 Interpretabilidade de modelos e SHAP

Para compreender o funcionamento interno do modelo RF e quantificar a influência das variáveis nas previsões, empregou-se a técnica SHAP. Essa abordagem é fundamentada na teoria de jogos cooperativos e atribui a cada preditor uma contribuição marginal para a saída do modelo, expressa por meio de valores de *Shapley* aproximados [Lundberg and Lee, 2017].

Essa contribuição média de uma variável para a previsão é calculada considerando todas as possíveis combinações de presença ou ausência dos demais preditores. Então, para cada instância, o método estima quanto a predição do modelo se altera quando um preditor é acrescentado a um subconjunto arbitrário de variáveis, e essa variação é então promediada sobre todas as combinações possíveis. Com isso, valores positivos indicam que a variável aumenta a estimativa em relação a uma referência, enquanto valores negativos indicam redução. Ao nível global, a média do valor absoluto dos valores *Shapley* sintetiza a importância relativa das variáveis, refletindo o quanto cada preditor altera sistematicamente as previsões, independentemente do sinal da contribuição.

No presente trabalho, o SHAP é empregado como ferramenta teórica de interpretabilidade para identificar quais atributos ambientais e meteorológicos exercem maior influência sobre as predições do RF. Dessa forma, a análise complementa as métricas tradicionais de desempenho ao fornecer transparência sobre o comportamento do modelo e sobre o papel relativo de cada variável nos resultados.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, apresentamos os principais estudos relacionados à previsão de internações por doenças respiratórias com base em dados ambientais e hospitalares utilizando modelos de aprendizado de máquina. A Seção 3.1 descreve a estratégia adotada para a seleção e priorização dos trabalhos revisados, incluindo os critérios de inclusão e exclusão. Em seguida, a Seção 3.2 discute os principais achados dos estudos priorizados, comparando as abordagens metodológicas, variáveis utilizadas e desempenho preditivo dos modelos. Além disso, são apresentados resumos individuais dos dez artigos com maior relevância para este TCC, seguidos pela Tabela 3, que sintetiza os elementos principais de cada estudo, como local de aplicação, melhores modelos utilizados e variáveis envolvidas. Por fim, a seção 3.3 apresenta uma análise crítica das convergências e divergências identificadas na literatura, destacando padrões metodológicos, lacunas existentes e implicações práticas para o contexto brasileiro.

3.1 Estratégia de Seleção e Priorização de Trabalhos Relacionados

A seleção dos artigos revisados neste trabalho foi guiada por um processo sistemático de busca e triagem, visando identificar publicações com alto alinhamento ao tema deste estudo. A construção da *string* de busca aplicada na base de dados *Scopus Elsevier* e foi construída com base na combinação de quatro grupos de palavras-chave principais, buscando contemplar artigos que tivessem tanto métodos computacionais quanto variáveis ambientais e de saúde pública. Esses quatro grupos incluem métodos de modelagem preditiva baseados em aprendizado de máquina, como RF e *Deep Learning*; poluentes atmosféricos relevantes, como $PM_{2.5}$ e dióxido de nitrogênio; desfechos de saúde ligados a doenças respiratórias; e o contexto urbano e de saúde pública. Foram também incluídos filtros negativos para excluir trabalhos que tratassem exclusivamente de câncer, dada a diferença de mecanismos patológicos envolvidos, além da limitação ao tipo documental “artigo de periódico” para garantir qualidade científica mínima.

A aplicação da *string* retornou 39 artigos, dos quais 22 foram considerados pertinentes após triagem baseada em título e resumo. Destes, 10 artigos foram priorizados para análise

aprofundada, por apresentarem forte relação com o tema e relevância metodológica, conforme descrito nas seções seguintes. Esse processo garante que a revisão bibliográfica se mantenha focada em trabalhos empíricos aplicáveis ao contexto da saúde pública urbana em países em desenvolvimento, como o Brasil. A *string* utilizada para a busca é apresentada na Listagem 3.1

Listing 3.1: Consulta utilizada na base *Scopus*

```
TITLE("machine␣learning" OR "random␣forest" OR "neural␣network" OR "deep
   ␣learning" OR "predictive␣model*")
AND TITLE-ABS-KEY("air␣quality" OR "air␣pollut*" OR pollutant* OR "
   particulate␣matter" OR
"nitrogen␣dioxide" OR "carbon␣monoxide" OR ozone OR "sulfur␣dioxide")
AND TITLE-ABS-KEY("respiratory␣disease*" OR "respiratory␣admission*" OR
"respiratory␣hospitalization*" OR "COPD")
AND TITLE-ABS-KEY("machine␣learning" OR "predictive␣model*" OR "forecast
   *" OR
"random␣forest" OR "neural␣network*" OR "deep␣learning" OR "regression␣
   model*" OR "time␣series")
AND TITLE-ABS-KEY("urban" OR "city" OR "population␣health" OR "
   metropolitan" OR "public␣health")
AND NOT TITLE-ABS-KEY(cancer OR cancers OR cancerous)
AND (LIMIT-TO(DOCTYPE, "ar"))
```

3.1.1 Critérios de inclusão

Os artigos selecionados foram incluídos com base nos seguintes critérios:

- Utilização de dados ambientais reais, obtidos de sensores ou estações oficiais.
- Presença de dados de saúde pública (internações, atendimentos ambulatoriais e mortalidade).
- Aplicação de métodos de aprendizado de máquina na modelagem ou predição dos desfechos.
- Enfoque empírico, com dados reais e metodologia detalhada.
- Publicação em periódico indexado e com replicabilidade metodológica.

3.1.2 Critérios de Exclusão

Foram excluídos artigos que apresentavam uma ou mais das seguintes características:

- O foco do artigo não envolvia doenças respiratórias (exemplo: somente câncer ou doenças cardiovasculares).
- Estudos com ênfase em ambientes *indoor* (escolas, escritórios) sem relação com exposição ambiental urbana.
- Artigos puramente metodológicos, sem aplicação empírica a dados reais.
- Trabalhos que não apresentavam dados de saúde pública (internações, atendimentos, mortalidade).

3.1.3 Critérios de Priorização

Após a triagem inicial, os trabalhos foram classificados quanto à relevância para o TCC, com base em sua contribuição metodológica e aderência ao tema. Eles foram divididos em três níveis de prioridade, incluindo:

- **Alta:** estudos que aplicam aprendizado de máquina para prever internações ou atendimentos respiratórios com dados ambientais e de saúde pública, em países em desenvolvimento como o Brasil e em contextos urbanos.
- **Média:** trabalhos que usam aprendizado de máquina e tratam de desfechos relevantes de saúde pública que não são focados diretamente em internações ou atendimentos, como mortalidade, e também tratam de outras doenças, junto as respiratórias. Também inclui estudos em países desenvolvidos.
- **Baixa:** aplicam aprendizado de máquina, mas os desfechos de saúde possuem escopo ou dados limitados, ou foco secundário.

3.1.4 Resultados

A triagem dos artigos utilizando a metodologia PRISMA [Page et al., 2021] pode ser visualizado na Figura 8. A busca utilizando a *string* calibrada, composta por termos relacionados ao escopo do tema, foi realizada em 11 de junho de 2025, resultando na identificação de 39 artigos.

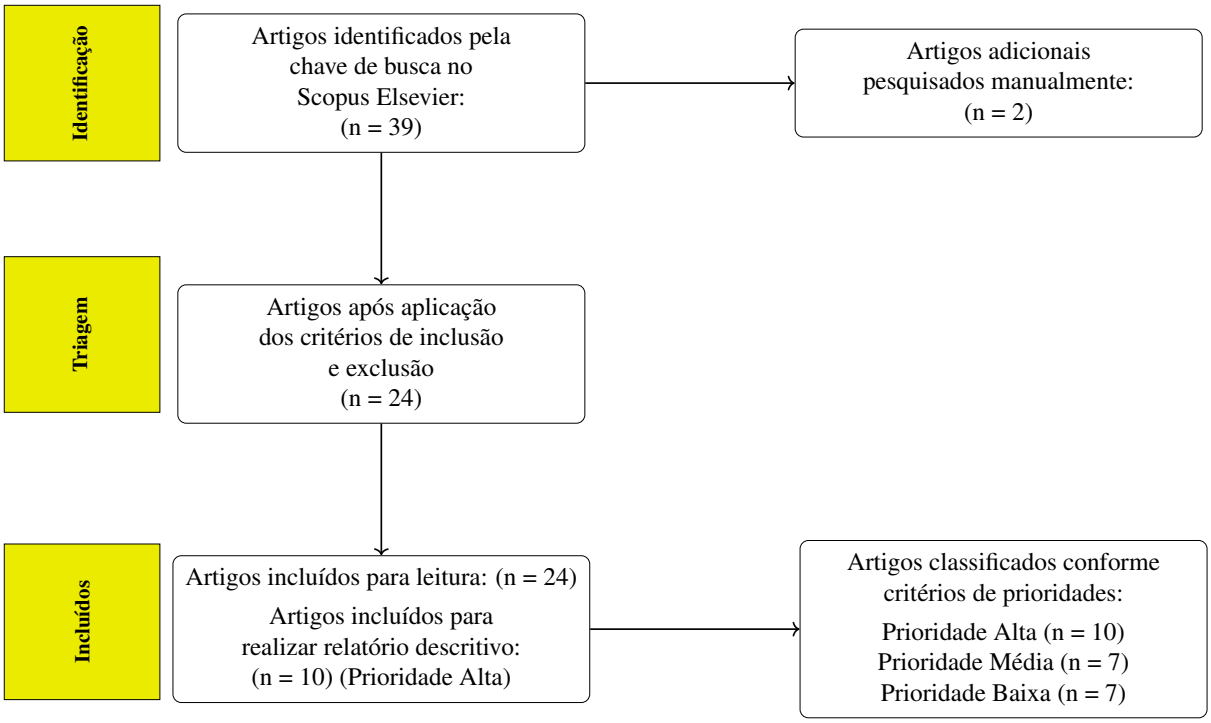

Figura 8: Fluxograma feito com PRISMA para seleção dos artigos.

3.2 Comparação entre os Trabalhos Relacionados

Nesta seção são comparados os dez estudos priorizados. A organização segue quatro eixos principais: modelos utilizados, variáveis empregadas, desfechos de saúde e estratégias de tratamento, otimização e interpretabilidade. A análise enfatiza não apenas o que cada artigo apresenta isoladamente, mas também como as abordagens dialogam entre si e revelam padrões comuns dentro do campo.

3.2.1 Modelos de Aprendizado de Máquina Utilizados

Uma comparação ampla entre os trabalhos revela um predomínio consistente de modelos não lineares, que se destacam pela capacidade de capturar relações complexas entre poluentes, condições meteorológicas e desfechos respiratórios. Entre eles, o RF ocupa posição central, apresentando desempenho superior em diversos cenários e aparecendo como melhor modelo em Reis et al. [2025], Temirbekov et al. [2023], Ji et al. [2023] e Yang et al. [2025]. Esse desempenho recorrente se explica por sua robustez diante de bases heterogêneas, pela tolerância a ruído e pela habilidade em modelar interações entre variáveis de diferentes naturezas. Redes neurais, sobretudo *Multilayer Perceptron* (MLP), *Extreme Learning Machines* (ELM) e *Radial*

Basis Function (RBF), também figuram com frequência na literatura, ganhando destaque em estudos brasileiros e chineses, especialmente quando combinadas em *ensembles* que ampliam o poder preditivo ao integrar múltiplas arquiteturas [Araujo et al., 2020; Kachba et al., 2020; Ji et al., 2023]. Já a *Support Vector Regression* (SVR) mostra vantagens específicas em contextos com séries temporais curtas ou com forte ênfase espacial, como observado em Yang et al. [2023] e Alvarez-Mendoza et al. [2020], nos quais a granularidade espacial elevada favorece seu mecanismo baseado em margens e *kernels*. Por outro lado, modelos *Long Short-Term Memory* (LSTM) aparecem de forma mais pontual, sendo escolhidos quando a dependência temporal é particularmente forte e a estrutura autorregressiva desempenha papel determinante na dinâmica do desfecho, como demonstrado em Lu et al. [2021]. Em contraste, modelos lineares como *Generalized Linear Model* (Modelo Linear Generalizado) (GLM), regressão múltipla ou *Ridge* são utilizados majoritariamente como *baselines*; em todos os estudos comparativos, são superados pelas alternativas não lineares, reforçando a natureza essencialmente não linear da relação entre poluição, clima e saúde respiratória.

3.2.2 Variáveis Ambientais e Meteorológicas Mais Relevantes

A comparação entre variáveis revela uma convergência expressiva entre os estudos revisados. Os poluentes atmosféricos regulados ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO e O_3) aparecem de maneira quase universal, refletindo tanto sua disponibilidade em redes de monitoramento quanto sua relevância epidemiológica consolidada [Reis et al., 2025; Ji et al., 2023; Temirbekov et al., 2023; Yang et al., 2023]. Entre esses, $PM_{2.5}$, PM_{10} e NO_2 emergem como os preditores mais influentes, enquanto O_3 apresenta com frequência associação negativa, modulada por padrões sazonais característicos de regiões temperadas. Em paralelo, variáveis meteorológicas como temperatura, umidade, vento e radiação se mostram essenciais para contextualizar a dinâmica dos poluentes e modular seus efeitos sobre o sistema respiratório. A inclusão dessas variáveis melhora consistentemente o desempenho dos modelos, sobretudo quando combinada a defasagens de poluentes e do próprio alvo. A importância das estruturas temporais também se reflete na adoção generalizada de *lags* entre 1 e 7 dias e de variáveis como mês ou estação, que capturam efeitos retardados e variações climáticas recorrentes [Reis et al., 2025; Yang et al., 2023; Araujo et al., 2020]. Um ponto de distinção aparece no estudo de Alvarez-Mendoza et al. [2020], que incorpora variáveis espaciais derivadas de sensoriamento remoto (*Normalized Difference Vegetation Index* (NDVI), *Land Surface Temperature* (LST) e bandas espectrais),

ampliando o escopo dos preditores e introduzindo uma dimensão geoespacial de alta resolução que não aparece nos demais trabalhos.

3.2.3 Desfechos de Saúde Considerados

A diversidade de desfechos considerados nos estudos não impede a identificação de um padrão dominante. A maioria dos trabalhos concentra-se em internações hospitalares, que constituem o desfecho mais disponível e mais diretamente associado aos picos de poluição, aparecendo em estudos nacionais e internacionais de diferentes escalas temporais e climáticas [Reis et al., 2025; Miranda et al., 2021; Araujo et al., 2020; Temirbekov et al., 2023]. Atendimentos ambulatoriais e de emergência surgem em menor volume, mas desempenham papel importante para entender respostas mais imediatas à deterioração da qualidade do ar, como discutido em Yang et al. [2025] e Lu et al. [2021]. Em contrapartida, estudos como Kachba et al. [2020] abordam morbidade e mortalidade, ampliando o escopo temporal e epidemiológico da análise ao considerar efeitos acumulados e tendências mais lentas. A abordagem espacial adotada por Alvarez-Mendoza et al. [2020] representa outra vertente relevante, ao trabalhar com dados distritais e modelar prevalência de doenças respiratórias em escala fina. Apesar dessas diferenças, permanece nítida a sensibilidade dos desfechos a $PM_{2.5}$, PM_{10} e NO_2, independentemente da forma de agregação temporal ou espacial, reforçando a robustez desses preditores.

3.2.4 Estratégias de Explicabilidade, Otimização e Tratamento dos Dados

Outra dimensão importante de comparação diz respeito às abordagens de interpretabilidade, otimização e tratamento de dados. Os estudos mais recentes demonstram uma incorporação crescente de técnicas de explicabilidade, com SHAP e *Local Interpretation through Model-agnostic Explanations* (Lógica Interpretável para Modelos Explicáveis) (LIME) aparecendo como ferramentas centrais para decompor a contribuição das variáveis e iluminar relações não lineares capturadas pelos modelos [Ji et al., 2023; Yang et al., 2025]. Em termos de otimização, métodos bioinspirados como *Particle Swarm Optimization* (PSO) são empregados para ajustar modelos lineares ou enriquecer *ensembles*, como visto em Araujo et al. [2020], evidenciando um interesse crescente em explorar estratégias híbridas. Já no tratamento de dados, observa-se grande valorização das defasagens temporais, normalização, ajustes sazonais e manejo de séries

incompletas, aspectos particularmente relevantes em países com monitoramento menos homogêneo, como o Brasil. No conjunto, essas estratégias revelam uma literatura que caminha em direção a *pipelines* mais integrados, com maior ênfase em interpretabilidade e rigor temporal, contribuindo para aproximar pesquisa aplicada e uso real em saúde pública.

3.3 Discussão

3.3.1 Síntese crítica das convergências

A revisão indica quatro princípios recorrentes: (i) modelos não lineares apresentam desempenho robusto em diferentes regimes de dados, com destaque para abordagens que conciliam preditores heterogêneos e interpretabilidade [Reis et al., 2025; Ji et al., 2023; Yang et al., 2025]; (ii) integração ambiental e meteorológica é indispensável, pois a variabilidade de poluentes e do clima modula consistentemente os desfechos respiratórios [Yang et al., 2023; Lu et al., 2021]; (iii) estrutura temporal (defasagens, sazonalidade e exposição acumulada) melhora a capacidade preditiva e a leitura epidemiológica dos efeitos retardados [Reis et al., 2025; Yang et al., 2023; Araujo et al., 2020]; e (iv) desfechos hospitalares predominam como alvo operacional, com sensibilidade robusta aos principais contaminantes [Kachba et al., 2020; Lu et al., 2021]. Em conjunto, esses elementos sustentam que o problema é tratável via aprendizado de máquina quando o desenho de modelo incorpora temporalidade e explicabilidade.

3.3.2 Divergências e diferenças metodológicas

As diferenças entre cidades, escalas e bases moldam o desempenho relativo dos modelos e a leitura causal dos preditores. Variações de granularidade temporal (diário vs. mensal), extensão das séries (curtas vs. longas) e regimes climáticos (temperados, tropicais, secos) produzem matrizes de correlação e padrões sazonais distintos, afetando a escolha de atributos, *detrending* e validação temporal [Yang et al., 2023; Lu et al., 2021]. Além disso, contrastes de cobertura de monitoramento e completude de dados de saúde, sendo mais estáveis em alguns contextos internacionais do que no Brasil, impõem decisões específicas de elegibilidade, imputação restrita e particionamento com *gap* [Temirbekov et al., 2023; Araujo et al., 2020; Kachba et al., 2020; Reis et al., 2025]. Portanto, não há um modelo universal, mas há protocolos que precisam ser contextualizados à disponibilidade e à qualidade local de dados, priorizando validação tempo-

ral, controle de sazonalidade e análise de viés.

3.3.3 Lacunas identificadas

Persistem lacunas que motivam avanços: (i) integração diária de múltiplos poluentes e meteorologia em estudos brasileiros, com documentação explícita de tratamento de ausências e consistência temporal [Araujo et al., 2020; Kachba et al., 2020; Miranda et al., 2021; Reis et al., 2025]; (ii) arquiteturas híbridas e *ensembles* com temporalidade, avaliadas de forma sistemática (combinação de dependência autorregressiva curta e efeitos acumulados ambientais) [Yang et al., 2025]; (iii) fontes adicionais de alta resolução, como sensoriamento remoto aplicado ao contexto urbano brasileiro com multiação e heterogeneidade espacial [Alvarez-Mendoza et al., 2020]; (iv) explicabilidade em operação, com protocolos padronizados para *Explainable artificial intelligence* (XAI) (SHAP, LIME, *Permutation Feature Importance* (PFI), *Partial Dependence Plot* (PDP)) sob sazonalidade e *drift* [Ji et al., 2023; Yang et al., 2025]; e (v) comparações multi-cidade com particionamento temporal comum e medidas de generalização fora da amostra [Araujo et al., 2020; Kachba et al., 2020; Reis et al., 2025].

3.3.4 Implicações para o trabalho proposto

As evidências orientam escolhas do TCC. Como modelo-base, adota-se RF com explicabilidade via SHAP, conciliando robustez a preditores heterogêneos e leitura operacional [Reis et al., 2025; Ji et al., 2023; Yang et al., 2025]. A engenharia temporal privilegia defasagens e médias móveis causais do alvo e dos poluentes, refletindo dependência de curto prazo e exposição acumulada [Lu et al., 2021; Yang et al., 2023]. Para sazonalidade, emprega-se codificação cíclica e STL, com *detrending* causal do alvo, separando variações de curto/médio prazo de mudanças lentas de regime e favorecendo calibração e R^2 em teste. A seleção de variáveis enfatiza NO_x/NO_2/NO, PM_{10}/$PM_{2.5}$ e O_3, em linha com associações robustas reportadas. A avaliação inclui um conjunto ampliado de métricas (MAE, RMSE, R^2, sMAPE, wMAPE e viés), respondendo ao chamado por diagnósticos mais informativos e úteis na gestão.

3.3.5 Considerações sobre o contexto brasileiro

O caso do Rio de Janeiro explicita desafios operacionais: monitoramento heterogêneo entre estações, com lacunas estruturais (p. ex., $PM_{2.5}$) que requerem regras claras de elegibilidade e imputação restrita em níveis horário/diário; bases de saúde do SUS com filtros clínico-administrativos e alinhamento geográfico à área de influência das estações; e heterogeneidade intraurbana (clima, relevo, tráfego, oferta de serviços) que demanda interpretação cuidadosa de risco no tempo e no espaço. Tais condições justificam um *pipeline* transparente (integração, tratamento de ausências, agregação diária, transformação de escala, engenharia de atributos) e validação temporal com *gap*, além de explicabilidade (SHAP, análise de viés por calendário) para apoio ao planejamento da rede pública.

3.3.6 Síntese final

Aprendizado de máquina se mostra adequado para previsão de desfechos respiratórios quando se combinam modelos não lineares, estrutura temporal e explicabilidade. As diferenças de contexto e de bases indicam que princípios de desenho (temporalidade, sazonalidade, validação e XAI) devem ser adaptados localmente. O presente trabalho incorpora esses princípios com ênfase em documentação de *pipeline*, *detrending* causal e avaliação ampliada, visando aplicação prática no município do Rio de Janeiro.

Artigo	Local	Melhor Modelo	Poluentes Usados	Variáveis Meteorológicas
Reis et al. [2025]	Maceió, Brasil	RF	SO_2, PM_{10}, O_3	Temperatura, umidade
Miranda et al. [2021]	São Paulo, Brasil	RBF Neural Network	$PM_{2.5}$, PM_{10}, CO, NO_2, O_3, SO_2	Não especificado
Yang et al. [2023]	Linyi, China	SVR	$PM_{2.5}$, PM_{10}, NO_2, CO	Temperatura, umidade, vento
Ji et al. [2023]	Taizhou, China	RF	$PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3	AQI, mês, poluente principal
Temirbekov et al. [2023]	Almaty, Cazaquistão	RF	PM_{10}, SO_2, NO_2, CO, metais	Não especificado
Yang et al. [2025]	Tianjin, China	RF + SHAP	$PM_{2.5}$, PM_{10}, SO_2, O_3, NO_2, CO	Temperatura, umidade, vento, radiação
Lu et al. [2021]	Pequim, China	LSTM	$PM_{2.5}$	Temperatura, umidade, vento
Alvarez-Mendoza et al. [2020]	Quito, Equador	SVR	$PM_{2.5}$, CO, SO_2, NO_2, O_3	NDVI, LST, radiação, umidade, temperatura
Kachba et al. [2020]	São Paulo, Brasil	ELM (morb.), ESN (mort.)	CO, NO_x, O_3, SO_2, PM_{10}	Sazonalidade, frota veicular
Araujo et al. [2020]	Campinas e São Paulo, Brasil	Ensemble (MLP)	PM_{10}	Temperatura, umidade, feriado
Este Trabalho	Rio de Janeiro, Brasil	RF	$PM_{2.5}$, PM_{10}, CO, NO, NO_2, NO_x, SO_2, O_3	Temperatura, umidade relativa do ar e precipitação pluviométrica

Tabela 3: Resumo comparativo dos estudos revisados sobre predição de doenças respiratórias, incluindo os modelos utilizados, poluentes considerados e variáveis meteorológicas analisadas.

Capítulo 4

Metodologia de Integração de Dados, Engenharia de Atributos e Modelagem Preditiva

Neste capítulo, descrevemos o fluxo metodológico utilizado para construir o conjunto de dados final e desenvolver um modelo preditivo para estimar o número de internações. Na Seção 4.1, apresentamos a formulação do problema e a justificativa do modelo escolhido. Na Seção 4.2, apresentamos uma visão geral do *pipeline* e das principais etapas do estudo. Na Seção 4.3, detalhamos a obtenção das fontes de dados (SIH/SUS e MonitorAr/DATA.RIO), bem como os critérios de seleção, filtragem e recorte temporal. Na Seção 4.4, descrevemos o pré-processamento das séries temporais, incluindo tratamento de ausências, consistência temporal, agregação para frequência diária e transformações estatísticas. Em seguida, na Seção 4.5, realizamos análises exploratórias para caracterizar a série de internações e as variáveis ambientais, investigando sazonalidade, tendência e dependência temporal. Na Seção 4.6, avaliamos as relações entre variáveis ambientais e internações por meio de correlações instantâneas e defasadas, além de medidas de exposição acumulada. Na Seção 4.7 é feita a construção da matriz final de *features*, reunindo atributos autorregressivos, sazonais, componentes STL e variáveis meteorológicas/poluentes. Por fim, na Seção 4.8, apresentamos o desenvolvimento do modelo preditivo, incluindo formulação da validação temporal, *detrending* causal, seleção de variáveis, ajuste de hiperparâmetros e métricas de avaliação.

4.1 Formulação do Problema e Justificativa do Modelo

Este estudo aborda um problema de previsão em séries temporais cujo objetivo é estimar o número diário de internações por doenças respiratórias no município do Rio de Janeiro. Definindo como y_t a variável-alvo no dia t (`num_internacoes`), busca-se aprender uma função preditiva $f(\cdot)$ capaz de produzir $\hat{y}_t$ a partir de informações disponíveis até $t-1$:

$$\hat{y}_t = f(X_t),$$

em que X_t reúne (i) atributos autorregressivos derivados do histórico recente das internações, (ii) atributos sazonais que representam ciclos anuais e semanais, e (iii) variáveis ambientais processadas (meteorologia e poluentes), incluindo defasagens e medidas de exposição acumulada construídas causalmente.

Por se tratar de um problema temporal, todas as etapas de preparação, treino e avaliação foram estruturadas para respeitar a ordem cronológica. A base final é organizada em frequência diária contínua, com previsão na granularidade de 1 dia. A avaliação adota um particionamento *hold-out* temporal, em que o conjunto de teste corresponde ao trecho mais recente da série, preservando a lógica de uso do modelo em cenário real. Além disso, foi utilizado um *gap* (em dias) entre treino e teste para mitigar dependência de curto prazo e evitar contaminação indireta causada por *features* construídas com janelas móveis e defasagens. Para a seleção de variáveis e o ajuste de hiperparâmetros, emprega-se validação cruzada temporal com *folds* ordenados no tempo (janelas expansivas ou deslizantes), de forma que cada avaliação simule o treinamento no passado e o teste em um bloco futuro.

O modelo preditivo adotado é o RF, um método de *ensemble* baseado em múltiplas árvores de decisão. A escolha do RF é motivada por sua capacidade de capturar relações não lineares e interações entre preditores sem exigir suposições fortes sobre distribuições, além de apresentar robustez relativa a distintas e a valores extremos, características frequentes em variáveis ambientais. Além disso, o RF é amplamente utilizado na literatura em aplicações que combinam poluição atmosférica, meteorologia e desfechos em saúde, com desempenho elevado em tarefas de previsão de hospitalizações respiratórias [Yang et al., 2025; Temirbekov et al., 2023; Reis et al., 2025].

4.2 Visão Geral do Fluxo Metodológico

Esta seção apresenta uma visão geral do fluxo metodológico adotado no estudo, desde a obtenção e integração das bases até o treinamento e a avaliação do modelo preditivo. O objetivo é explicitar, resumidamente, as principais fases do *pipeline* que transforma dados ambientais e de saúde em uma base diária unificada, adequada tanto para análises exploratórias quanto para modelagem.

A solução proposta integra duas fontes de dados complementares. A primeira é o *MonitorAr* (DATARIO), que disponibiliza medições horárias de poluentes atmosféricos e variáveis meteorológicas em estações distribuídas pelo município do Rio de Janeiro. A segunda é o SIH/SUS

(via *PySUS*), composta por registros de internações hospitalares, a partir dos quais foi derivada a série diária de internações por doenças respiratórias. Em seguida, ambas as fontes são convertidas em uma mesma escala temporal (diária) e combinadas por alinhamento de datas.

Em termos gerais, o *pipeline* pode ser resumido nos seguintes passos:

1. Coleta e organização dos dados: obtenção das bases (MonitorAr e SIH/SUS) e padronização inicial de variáveis e formatos.

2. Unificação temporal: transformação das medições horárias ambientais para a escala diária e agregação dos registros de saúde por data.

3. Pré-processamento: limpeza e tratamento de inconsistências (valores ausentes, outliers e ajustes necessários), preparando as séries para análises e modelagem.

4. Análises exploratórias: investigação de dependência temporal/sazonalidade nas internações e análise de correlações entre variáveis ambientais e o desfecho, incluindo efeitos defasados quando aplicável.

5. Engenharia de atributos: criação e sistematização de *features* (por exemplo, variáveis defasadas, agregações móveis e componentes sazonais), formando a matriz final de entrada do modelo.

6. Divisão em treino e teste: separação temporal do conjunto para estimação dos parâmetros e avaliação fora da amostra.

7. Modelagem e avaliação: treinamento do modelo preditivo (RF) com os dados de treino e avaliação no conjunto de teste, produzindo previsões e métricas de desempenho.

A Figura 9 apresenta um diagrama resumindo as principais fases do fluxo.

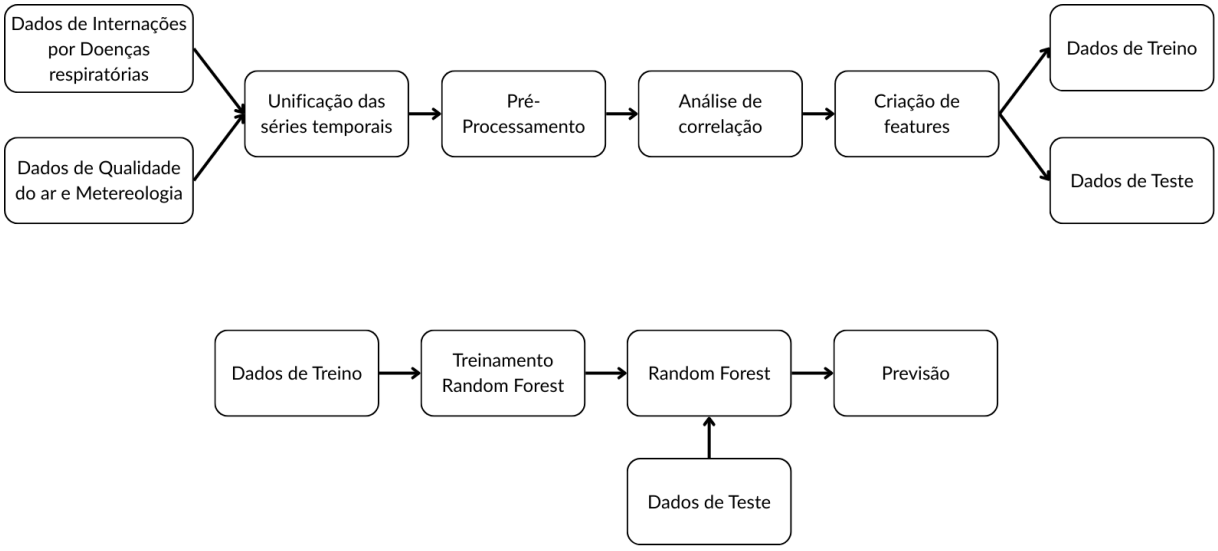

Figura 9: Diagrama do *pipeline* metodológico, destacando as etapas de obtenção, unificação e pré-processamento dos dados, análises exploratórias, engenharia de atributos e treinamento/avaliação do modelo preditivo.

4.3 Obtenção e Integração das Bases de Dados

Esta seção descreve a obtenção das duas fontes de dados utilizadas no estudo, internações hospitalares por doenças respiratórias (SIH/SUS) e medições ambientais e meteorológicas (MonitorAr / DATARIO), bem como os critérios de seleção, filtragem e recorte temporal adotados. Por fim, será detalhado o procedimento de integração das fontes e a construção da base diária unificada que serve de entrada para as análises exploratórias e para a modelagem preditiva nas seções subsequentes.

4.3.1 Dados de Internações Hospitalares

Os dados de internações por doenças respiratórias foram obtidos a partir do Sistema de Informações Hospitalares do SUS (SIH), disponível no portal do DATASUS (`https://datasus.saude.gov.br/`). A extração e o processamento inicial foram realizados em *Python* utilizando a biblioteca PySUS [Coelho et al., 2021], que permite o *download* automático dos arquivos do SIH, sua leitura em formato tabular e o salvamento em arquivos *Comma-Separated Values* (CSV).

Foram utilizados registros mensais do tipo Registro de Internação Hospitalar (RD), associados às Autorização de Internação Hospitalar (AIH) regulares, cobrindo o período de 2012 a 2024, com Unidade Federativa restrita ao estado do Rio de Janeiro (`uf="RJ"`). Os arquivos

foram obtidos mês a mês e consolidados em *dataframes* (*pandas*) para posterior filtragem e agregação temporal. A escolha deste intervalo visa manter a coerência temporal com os dados dos sensores obtidos do MonitorAr.

A variável temporal utilizada como referência foi `DT_INTER`, originalmente codificada no formato numérico `AAAAMMDD`. Essa coluna foi convertida para o tipo data (formato `AAAA-MM-DD`) e utilizada para construir a série diária de internações após a aplicação dos critérios clínicos e geográficos descritos na Subseção 4.3.3.

4.3.2 Dados de Qualidade do Ar e Meteorologia

Os dados de qualidade do ar foram obtidos por meio do conjunto MonitorAr, disponibilizado publicamente pelo DATA.RIO [Prefeitura do Rio de Janeiro, 2024b], o portal oficial de dados abertos da Prefeitura da Cidade do Rio de Janeiro (`https://www.data.rio/datasets/PCRJ::dados-hor%C3%A1rios-do-monitoramento-da-qualidade-do-ar-monitorar/about`). Lançado em 2017 como sucessor do *Armazém de Dados*, o DATA.RIO tem como objetivo centralizar, integrar e disponibilizar dados municipais com foco em transparência e formulação de políticas públicas baseadas em evidências.

O programa MonitorAr realiza o monitoramento da qualidade do ar por meio de oito estações distribuídas pelas zonas Norte, Sul, Oeste e Central do município do Rio de Janeiro: Bangu, Campo Grande, Pedra de Guaratiba, Irajá, Tijuca, São Cristóvão, Copacabana e Centro. Essas estações registram medições horárias de material particulado fino ($PM_{2.5}$, PM_{10}), óxidos de nitrogênio (NO, NO_2, NO_x), dióxido de enxofre (SO_2), ozônio (O_3) e monóxido de carbono (CO), sendo que as concentrações de CO são expressas em partes por milhão (ppm), enquanto as concentrações dos demais poluentes são medidas em microgramas por metro cúbico de ar ($\mu g/m^3$). Adicionalmente, são monitoradas variáveis meteorológicas utilizadas neste estudo, tais como temperatura do ar ($^\circ$C), umidade relativa do ar (%) e precipitação pluviométrica (mm).

A Tabela 4 apresenta a relação entre as estações de monitoramento e os poluentes medidos em cada uma:

As estações de monitoramento possuem diferentes datas de início de operação no período analisado. As estações Centro, Copacabana, São Cristóvão e Tijuca iniciaram as medições em setembro de 2011, enquanto as estações de Bangu, Campo Grande, Pedra de Guaratiba e Irajá passaram a operar a partir de setembro de 2012. As observações são realizadas em frequência

Tabela 4: Estações de monitoramento do programa MonitorAr e poluentes monitorados.

Estação	**Poluentes Monitorados**
Centro (CA)	O_3, CO, PM_{10}
Copacabana (AV)	SO_2, O_3, CO, PM_{10}
São Cristóvão (SC)	SO_2, O_3, CO, PM_{10}
Tijuca (SP)	SO_2, NO_x, O_3, CO, PM_{10}
Irajá (IR)	SO_2, NO_x, O_3, CO, HC, $PM_{2.5}$, PM_{10}
Bangu (BG)	SO_2, NO_x, O_3, CO, HC, PM_{10}
Campo Grande (CG)	SO_2, NO_x, O_3, CO, HC, PM_{10}
Pedra de Guaratiba (PG)	O_3, PM_{10}

horária, totalizando até 24 registros por dia para cada variável monitorada. Ressalta-se que, até o momento deste estudo, os dados consolidados e disponíveis para análise abrangiam apenas o período até o final do ano de 2024.

Apesar da frequência horária e da longa duração das séries temporais, as bases apresentam proporções variáveis de valores ausentes entre estações e variáveis. De forma geral, as variáveis meteorológicas (`chuva`, `temp` e `ur`) apresentaram baixas taxas de ausência, com percentuais inferiores a 10% na maioria das estações.

Por outro lado, variáveis relacionadas à qualidade do ar, como `co`, `no`, `no2`, `nox`, `so2` e `pm2_5`, apresentaram lacunas significativas. Em algumas estações, como Pedra de Guaratiba, Centro e Copacabana, determinados poluentes (ex.: `co`, `no` e `pm2_5`) estiveram ausentes em 100% dos registros, o que é esperado, visto que essas estações não medem essas variáveis.

A variável `pm2_5`, por exemplo, apresentou ausência total de dados nas estações de Bangu, Campo Grande, Centro, Copacabana, Pedra de Guaratiba, São Cristóvão e Tijuca, existindo apenas na estação de Irajá.

Para este estudo, foram utilizados os dados referentes ao período de 2012 a 2024, abrangendo os poluentes atmosféricos e as variáveis climáticas coletadas nas oito estações mencionadas.

Além disso, foram incorporados os dados do AQI, também disponibilizados pelo portal DATARIO. O AQI fornece uma classificação padronizada da qualidade do ar com base nas concentrações de poluentes: quanto maior for o índice, pior a qualidade do ar [United States Environmental Protection Agency, nd]. Neste trabalho, o AQI foi tratado como uma variável ambiental adicional, de forma que oito séries foram consideradas na modelagem: $PM_{2.5}$, PM_{10},

CO, NO_2, NO_x, O_3, SO_2 e o próprio AQI.

A escolha desses poluentes e do AQI baseou-se em evidências da literatura que apontam sua associação com desfechos de saúde respiratória, conforme discutido por Alvarez-Mendoza et al. [2020]; Ji et al. [2023]; Miranda et al. [2021], bem como nas diretrizes da EPA e da Organização Mundial da Saúde (OMS).

4.3.3 Critérios de Seleção, Filtragem e Recorte Temporal.

Os dados ambientais, foram selecionadas as medições horárias disponíveis no MonitorAr para as estações listadas na Subseção 4.3.2. Considerando que diferentes estações monitoram diferentes poluentes e que há ausência estrutural em algumas variáveis, a inclusão efetiva de cada variável na base final depende da disponibilidade mínima de dados após os procedimentos de tratamento e agregação descritos na Seção 4.4.

Já a seleção dos registros de saúde seguiu critérios clínicos e administrativos, visando restringir a análise a internações por doenças do aparelho respiratório. Para isso, são aplicados os seguintes critérios de inclusão e exclusão:

- Tipo de registro: seleção apenas de registros do tipo RD, correspondentes às AIH regulares.
- Diagnóstico principal: filtragem dos registros nos quais a variável `DIAG_PRINC` (diagnóstico principal) inicia-se com a letra "J", conforme a CID, que agrupa as doenças do aparelho respiratório nos códigos J00–J99.
- Identificador do registro: manutenção apenas dos registros com `IDENT = '1'`, para considerar unicamente as internações principais registradas no sistema.
- Período de estudo: restrição da amostra ao intervalo de 01/01/2012 a 31/12/2024, para seguir o mesmo intervalo temporal dos sensores

4.3.4 Integração das Fontes e Construção da Base Diária Unificada

A integração das fontes foi realizada por meio da construção de uma base diária unificada, com uma linha por dia e colunas correspondentes às variáveis explicativas (ambientais e me-

Tabela 5: Exemplo do formato final dos dados de internações por doenças respiratórias após a obtenção

data_formatada	MUNIC_RES	SEXO	DT_INTER	DT_SAIDA	DIAG_PRINC	IDADE	DIAS_PERM	MORTE	CID_MORTE
2012-01-01	330350	1	20120101	20120114	J189	56	13	1	J960
2012-01-01	330250	3	20120101	20120110	J188	80	9	1	J188
2012-01-01	330455	1	20120101	20120103	J351	8	2	0	
2012-01-01	330455	1	20120101	20120104	J353	8	3	0	

teorológicas) e à variável-alvo (internações respiratórias). O processo ocorreu em três etapas principais:

1. Agregação diária das séries ambientais: as medições horárias do MonitorAr foram convertidas para a escala diária por estação, por meio de agregações apropriadas (por exemplo, médias diárias para concentrações e temperatura, e acumulado diário para precipitação, quando aplicável). Em seguida, obteve-se uma série municipal por variável a partir da agregação entre estações disponíveis em cada dia (por exemplo, média entre estações), mantendo cada variável separadamente.

2. Agregação diária das internações: após a filtragem clínica e administrativa, os registros do SIH/SUS foram agregados por data de internação (`DT_INTER`), resultando no total diário de internações por doenças respiratórias no recorte definido.

3. Alinhamento temporal e junção das fontes: as séries diárias (ambientais e de internações) foram alinhadas pela data e combinadas em um único *dataset*. Assim, cada observação diária contém como preditores as condições ambientais do dia (e, nas etapas posteriores, versões defasadas e transformadas dessas variáveis) e, como resposta, o número de internações respiratórias.

A base diária unificada construída nesta seção constitui o insumo central para as análises de sazonalidade, dependência temporal e correlação apresentadas nas próximas seções, além de fundamentar a etapa de engenharia de atributos e a modelagem preditiva descritas adiante.

4.4 Pré-processamento das Séries Temporais

Esta seção descreve os procedimentos de pré-processamento aplicados às séries temporais das duas fontes do estudo. O objetivo foi garantir consistência temporal, reduzir efeitos de ausência e ruídos nas medições e padronizar ambas as fontes para uma representação diária

compatível, formando entradas adequadas para as análises estatísticas e para a modelagem preditiva.

4.4.1 Tratamento de Ausências e Consistência Temporal

Dados ambientais (estações): Os dados do MonitorAr são registrados em frequência horária, com até 24 observações por dia para cada variável e estação. Como a etapa seguinte do estudo utiliza uma base diária, o tratamento de ausências foi conduzido prioritariamente no nível horário, de modo a recuperar lacunas curtas antes do cálculo dos agregados diários.

Inicialmente, o conjunto original foi segmentado por estação, permitindo inspeção e limpeza individualizadas. Em seguida, para cada variável e para cada dia, foi calculado a proporção de medições ausentes. Adotou-se uma regra conservadora, onde somente os dias com até 6 horas faltantes (no máximo 25% das observações do dia) foram considerados elegíveis para imputação. Dias com mais de 6 horas ausentes para determinada variável permaneceram como faltantes, evitando que o valor diário fosse calculado a partir de uma fração pequena do dia e, consequentemente, pouco representativa.

Nos dias elegíveis, as ausências foram preenchidas por interpolação linear no eixo temporal horário. Quando as lacunas ocorriam nos extremos do dia (primeiras ou últimas horas), aplicou-se preenchimento pelo valor válido mais próximo no próprio dia (*forward/backward fill* restrito ao dia), reduzindo o risco de propagação de informação entre dias distintos.

Dados de internações: O tratamento dos dados de internação hospitalar teve como objetivo gerar uma série temporal diária com o número de internações por doenças respiratórias no município do Rio de Janeiro, compatível com os dados dos sensores tratados previamente.

Inicialmente, foram concatenados todos os arquivos CSV gerados na etapa de obtenção dos dados de (Seção 4.3), resultando em uma base única com **5.002.306** registros de internações no período de 2012 a 2024. A partir disso, foram aplicados filtros clínicos, geográficos e de consistência para delimitar a registros de interesse e garantir maior qualidade aos dados utilizados na modelagem.

O primeiro critério de seleção baseou-se na variável `DIAG_PRINC`. Foram mantidos somente os registros cujos códigos da CID-10 pertencem ao subconjunto de doenças, como pneumonias (`J15`, `J18`), bronquiolite aguda (`J21`), insuficiência respiratória (`J96`), bronquites (`J40`, `J41`, `J42`), doença pulmonar obstrutiva crônica/enfisema (`J43`, `J44`) e asma (`J45`, `J46`). A es-

colha dessas doenças foi baseada em evidências da literatura que apontam associações entre essas condições e variáveis ambientais, como poluentes atmosféricos e condições meteorológicas [Assunção et al., 2023].

Em seguida, aplicou-se um filtro geográfico na coluna `MUNIC_RES`, que identifica o município de residência do paciente. Foram selecionados somente os registros referentes ao município do Rio de Janeiro (`330455`) e a municípios vizinhos da Região Metropolitana (`330510` – São João de Meriti, `330320` – Nilópolis, `330285` – Mesquita, `330045` – Belford Roxo, `330350` – Nova Iguaçu, `330170` – Duque de Caxias). Essa filtragem foi feita para alinhar a área de captação das internações com a área de influência das estações de monitoramento de qualidade do ar.

Além disso, foi feito um filtro na variável `DIAS_PERM`, que indica o número de dias de permanência hospitalar. Foram mantidos apenas os registros em que `DIAS_PERM` fosse maior que zero, de modo a excluir registros possivelmente administrativos, cancelamentos ou internações com informação de permanência inconsistente.

Após a aplicação desses filtros, a base passou de **5.002.306** para **1.412.369** registros de internações, correspondendo a uma redução de aproximadamente **71,8%** em relação ao conjunto original, preservando cerca de **28,2%** dos registros considerados mais aderentes ao escopo epidemiológico do estudo.

Na etapa seguinte, as internações foram agregadas por data utilizando a variável `data_formatada` (derivada da `DT_INTER` e expressa no formato `AAAA-MM-DD`), contabilizando-se, para cada dia, o número total de internações por doenças respiratórias. Essa agregação resultou em uma série temporal diária representada pela variável `num_internacoes`. A inspeção exploratória dessa série indicou que, no período analisado, não foram observados dias com contagem igual a zero, isto é, houve pelo menos uma internação respiratória em todos os dias do intervalo analisado, por conta disso não houve a necessidade de tratar valores ausentes.

Preditores: Embora parte do tratamento de ausências tenha sido realizada ainda no nível horário das estações, constatou-se que o conjunto final de preditores ainda apresentava valores faltantes, seja por limitações estruturais (como estações que não medem certos poluentes), seja por lacunas adicionais de medição. Por isso, o tratamento final das ausências foi realizado apenas após a unificação dos *datasets*, quando todas as variáveis preditoras já se encontravam alinhadas no mesmo índice temporal e reunidas em um único *dataframe*.

Nesse estágio final, para lidar com ausências nos preditores sem introduzir vazamento de

Tabela 6: Exemplo da série temporal diária construída a partir dos registros do SIH/SUS após a aplicação dos filtros clínicos e geográficos e a agregação por data de internação (`DT_INTER`). Cada linha representa um dia do período analisado e a variável `num_internacoes` corresponde à contagem total de internações respiratórias naquele dia, servindo como variável-alvo (y_t) nas etapas de análise temporal e modelagem preditiva.

data_formatada	num_internacoes
2012-01-01	507
2012-01-02	728
2012-01-03	429
2012-01-04	416
2012-01-05	403
2012-01-06	520
2012-01-07	299
2012-01-08	390
2012-01-09	442
2012-01-10	546

informação, adotou-se uma estratégia em duas etapas:

1. Criação de indicadores de ausência: para cada preditor x_j, foi criada uma variável binária `miss_`x_j, assumindo valor 1 quando x_j estava ausente no dia t, e 0 caso contrário. Essas *flags* permitem informar explicitamente ao modelo quando um valor foi imputado, preservando o sinal de que ocorreu indisponibilidade de medição.

2. Imputação por *forward-fill* limitado: aplicou-se *forward-fill* apenas nas variáveis preditoras, preenchendo valores ausentes com o valor válido mais recente no passado. Esse preenchimento foi realizado limitadamente (sem propagação por longos períodos), reduzindo o risco de gerar sequências artificiais extensas. Importante: não foi realizado preenchimento no alvo (`num_internacoes`), mantendo a série de internações inalterada.

Esse procedimento permitiu manter o conjunto temporal consistente e utilizável para treino e teste, ao mesmo tempo em que preservou informação sobre a qualidade/ disponibilidade das medições por meio das variáveis indicadoras de *missing*.

4.4.2 Agregação para Frequência Diária

Após preencher os dados de cada estação, foi feita uma agregação para a escala diária. Para as variáveis meteorológicas e poluentes, o valor diário foi calculado como a média das 24 observações do dia. A variável `chuva`, por representar um volume acumulado, foi convertida

para a escala diária por meio da soma das medições horárias. Essa agregação gerou um *dataset* diário por estação, salvo em arquivo CSV (um arquivo por estação).

Por fim, os *datasets* diários das oito estações foram reunidos em um único *dataframe*. Em seguida, para obter uma série temporal em cada data, foi realizada uma agregação por dia em que, para cada variável, calculou-se a média entre as estações disponíveis naquele dia. Assim, cada linha corresponde a um dia, e cada coluna corresponde ao valor diário médio de uma variável específica na cidade (por exemplo, `temp`, `ur`, `o3`, `pm10`, etc.), e não a uma média única combinando todos os poluentes. Esse conjunto consolidado foi então utilizado como base para a etapa de transformação estatística.

Tabela 7: Exemplo do formato final do *dataset* diário de variáveis meteorológicas e poluentes (média diária; chuva agregada por soma diária).

data_formatada	chuva	temp	ur	co	no	no2	nox	so2	o3	pm10	pm2_5
2012-01-01	12,250	25,835	92,165	0,426	3,613	23,520	27,130	2,673	23,059	23,925	14,619
2012-01-02	56,050	22,836	95,589	0,305	12,675	27,160	39,842	1,794	20,136	13,872	5,083
2012-01-03	0,025	24,948	76,139	0,260	17,175	28,731	45,883	3,919	15,718	24,064	4,208
2012-01-04	0,050	26,006	72,904	0,275	24,746	40,338	65,049	3,124	25,002	35,773	15,729
2012-01-05	0,000	26,498	75,515	0,271	16,643	34,914	51,559	3,066	33,646	32,901	10,917

4.4.3 Transformações Estatísticas (Yeo–Johnson)

Séries de poluentes atmosféricos frequentemente apresentam assimetria, caudas longas e valores extremos, o que pode aumentar a influência de outliers e dificultar análises estatísticas baseadas em relações aproximadamente lineares. Para diminuir esses efeitos e estabilizar a variância, aplicou-se uma etapa de transformação de potência nas variáveis contínuas selecionadas.

Antes da aplicação de qualquer técnica de transformação, foi conduzida uma análise exploratória da distribuição de cada variável contínua. Para isso, foram utilizados histogramas, gráficos Q-Q (*quantile-quantile plot*) e o teste de normalidade de Kolmogorov-Smirnov, de modo a identificar desvios significativos da normalidade.

Observou-se que a maioria das variáveis apresentava distribuições assimétricas e/ou com caudas pesadas, com exceção da temperatura e umidade relativa o ar. Esse comportamento tende a aumentar a influência de valores extremos e tornar a variância instável ao longo do tempo, o que pode dificultar análises estatísticas baseadas em relações aproximadamente lineares e reduzir a interpretabilidade de diagnósticos exploratórios.

Como exemplo, a variável monóxido de nitrogênio (NO) apresentou distribuição visivelmente assimétrica, com cauda à direita (positiva) e forte desvio da normalidade, conforme é mostrado pelos gráficos e testes a seguir.

O histograma da Figura 10 demonstra uma assimetria à direita acentuada. A média (13,95 μg/m^3) é significativamente maior que a mediana (11,03 μg/m^3), e o desvio padrão elevado (10,16 μg/m^3) reforça a presença de *outliers*.

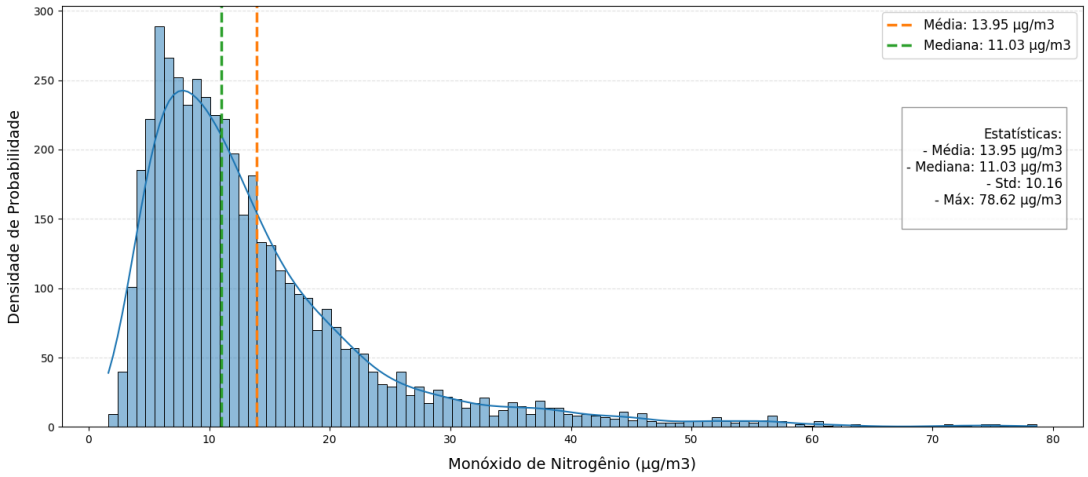

Figura 10: Distribuição da variável Monóxido de Nitrogênio (NO) antes da transformação. A linha laranja representa a média e a linha verde, a mediana, permitindo observar a assimetria da distribuição.

A Figura 11 ilustra o Q-Q Plot da variável NO, em que se observa um desvio acentuado dos pontos em relação à linha vermelha de referência (linha de normalidade teórica), caracterizando uma distribuição não-normal.

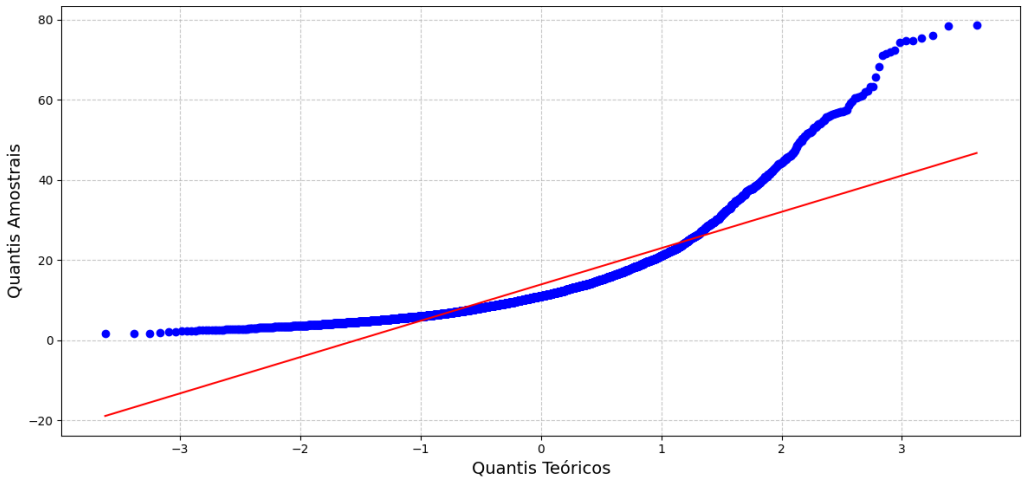

Figura 11: Q-Q Plot da variável NO antes da transformação, mostrando forte desvio da normalidade, especialmente nas caudas. Esse padrão justifica a necessidade de transformação dos dados.

Para confirmar estatisticamente essa observação, foi aplicado o teste de normalidade de Kolmogorov-Smirnov, no qual resultou em uma estatística W = 0,168 e um valor de p extremamente baixo ($p < 0{,}00001$), mais precisamente $p = 3{,}04 \times 10^{-25}$. Como esse valor é inferior ao limiar de significância usual ($\alpha = 0{,}05$), rejeita-se a hipótese nula de que a amostra segue uma distribuição normal. Portanto, a variável `no` não apresenta distribuição normal.

Diante da evidência de não-normalidade e da presença de assimetria, optou-se pela aplicação de uma transformação de potência do tipo **Yeo–Johnson**. Essa transformação visa reduzir a assimetria e estabilizar a variância, produzindo distribuições mais regulares para análises estatísticas subsequentes.

A transformação foi aplicada utilizando a classe `PowerTransformer` da biblioteca *scikit-learn*, com `method="yeo-johnson"` e `standardize=False`. O parâmetro de potência (λ) foi estimado automaticamente a partir dos próprios dados durante o ajuste. Como o transformador não aceita valores ausentes, o ajuste e a transformação foram realizados apenas sobre as observações não nulas, preservando `NaN` nas posições originalmente ausentes.

Após a aplicação da transformação, foi feita uma nova rodada de verificação da distribuição, com os mesmos métodos utilizados anteriormente, de modo a validar a efetividade da transformação.

A Figura 12 e a Figura 13 ilustram a variável NO após a transformação, evidenciando melhora na simetria da distribuição e alinhamento com a normalidade.

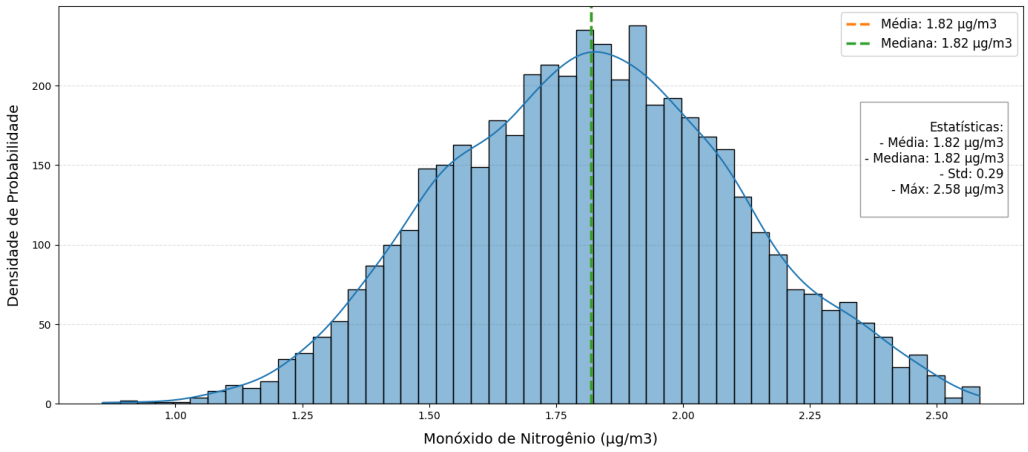

Figura 12: Distribuição da variável Monóxido de Nitrogênio (NO) após aplicação da transformação Yeo-Johnson. A distribuição resultante mostra simetria e sobreposição com a curva normal ajustada, indicando que a transformação foi eficaz em aproximar os dados de uma distribuição normal.

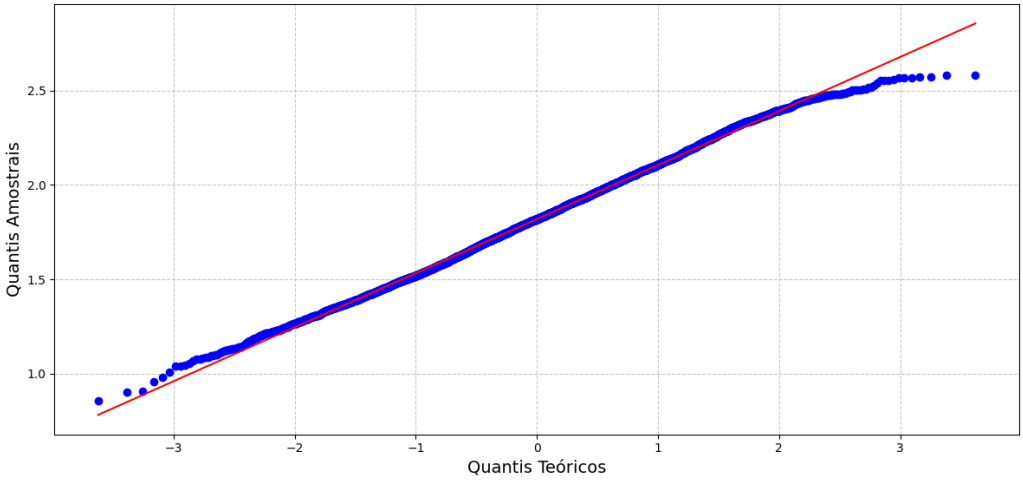

Figura 13: Q-Q Plot da variável NO após a transformação Yeo-Johnson, evidenciando um bom alinhamento com a linha de referência (vermelha). Esse comportamento indica que os dados transformados seguem aproximadamente uma distribuição normal.

Essa normalização foi confirmada pelo novo teste de Kolmogorov-Smirnov foi aplicado à variável transformada, que retornou uma estatística W = 0,025 e valor de p = 0,548, indicando que, após a transformação, não há evidências para rejeitar a hipótese de normalidade (p > 0,05).

Tabela 8: Exemplo do formato do *dataset* diário com as variáveis meteorológicas e poluentes após o pré-processamento e a transformação de potência (Yeo–Johnson). Cada linha representa um dia e cada coluna corresponde a uma variável ambiental agregada para a escala diária

data_formatada	chuva	temp	ur	co	no	no2	nox	so2	o3	pm10	pm2_5
2012-01-01	0,85	25,84	92,17	0,22	1,26	4,18	2,77	1,57	6,47	3,02	3,84
2012-01-02	0,89	22,84	95,59	0,18	1,88	4,42	3,02	1,19	6,01	2,56	2,24
2012-01-03	0,02	24,95	76,14	0,17	2,02	4,51	3,11	2,00	5,25	3,02	2,01
2012-01-04	0,05	26,01	72,90	0,17	2,18	5,10	3,32	1,74	6,75	3,35	3,97
2012-01-05	0,00	26,50	75,51	0,17	2,01	4,84	3,18	1,72	7,89	3,28	3,34

4.4.4 Unificação das Bases e Construção do *Dataset* Final

Após a obtenção da série diária de internações (Tabela 6) e da série diária municipal de variáveis meteorológicas e poluentes (Tabela 8), realizou-se a unificação das bases para formar o *dataset* final utilizado nas análises de correlação e na modelagem preditiva.

Como etapa inicial, foi feita uma integração por alinhamento temporal, isto é, por meio de uma junção (*merge*) utilizando a data como chave.

O *dataset* unificado contém, portanto, uma linha por dia e colunas que incluem (i) a variável-alvo `num_internacoes` e (ii) as variáveis explicativas ambientais e meteorológicas. Esse con-

junto foi inicialmente utilizado para analisar a série temporal (Seção 4.5) e, em seguida, serviu de base para as análises de correlação entre qualidade do ar e internações (Seção 4.6), nas quais são investigadas possíveis relações diretas e defasadas entre exposição ambiental e desfechos respiratórios.

4.5 Análises Temporais Exploratórias

Esta seção apresenta as análises temporais exploratórias realizadas sobre a série diária gerada. O objetivo é caracterizar o comportamento do alvo ao longo do tempo, identificando padrões de sazonalidade (anual e semanal), tendência de longo prazo, dependência temporal de curto prazo e analisar a distribuição das variáveis ambientais. Esses achados fundamentam as escolhas de *features* autorregressivas e sazonais utilizadas na etapa de engenharia de atributos (Seção 4.7).

4.5.1 Análise da Série de Internações

A Figura 14 apresenta a série temporal do número de internações no período de 2012 a 2024. A linha azul corresponde à contagem diária de internações, a linha laranja representa a média móvel de 7 dias e a linha tracejada indica a média global do período (255,9 internações/dia).

É possível ver uma variabilidade a curto prazo, com picos recorrentes ao longo dos anos. Esses picos ocorrem em intervalos aproximadamente anuais, sugerindo um padrão sazonal, compatível com a maior ocorrência de doenças respiratórias em determinados períodos do ano. Também é possível identificar uma tendência de queda gradual no nível médio de internações ao longo do tempo, especialmente a partir dos anos mais recentes da série.

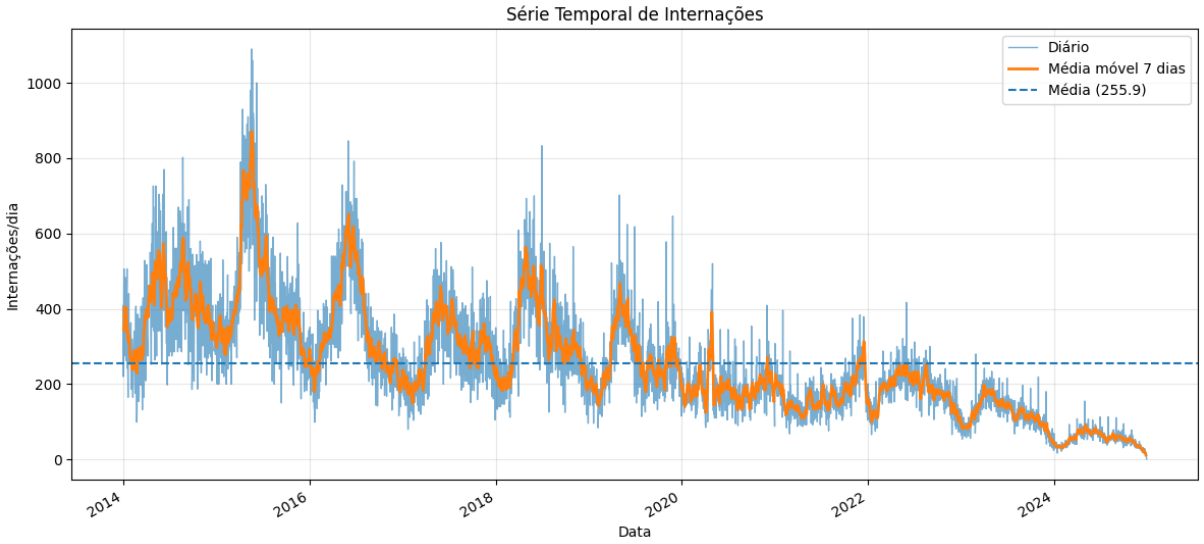

Figura 14: Série temporal diária das internações por doenças respiratórias, com a contagem diária (linha azul), a média móvel de 7 dias (linha laranja) e a média global do período (linha tracejada).

Para tentar encontrar algum padrão sazonal anual das internações, a série foi agregada por mês, calculando-se a média diária de internações em cada mês ao longo de todo o período. Dessa forma, cada barra do gráfico representa o número médio de internações por dia em um determinado mês, considerando todos os anos.

A Figura 15 mostra o perfil mensal. Observa-se um aumento constante das internações a partir de março, com valores mais altos entre abril e julho, quando a média diária ultrapassa 370 internações e atinge o pico em maio (cerca de 420 internações/dia). A partir de agosto, verifica-se uma queda gradual, com níveis mais baixos nos meses de primavera e início do verão (outubro a dezembro), em torno de 230–260 internações/dia. Esse comportamento indica um padrão sazonal claro, com maior carga de internações por doenças respiratórias concentrada nos meses de outono e inverno.

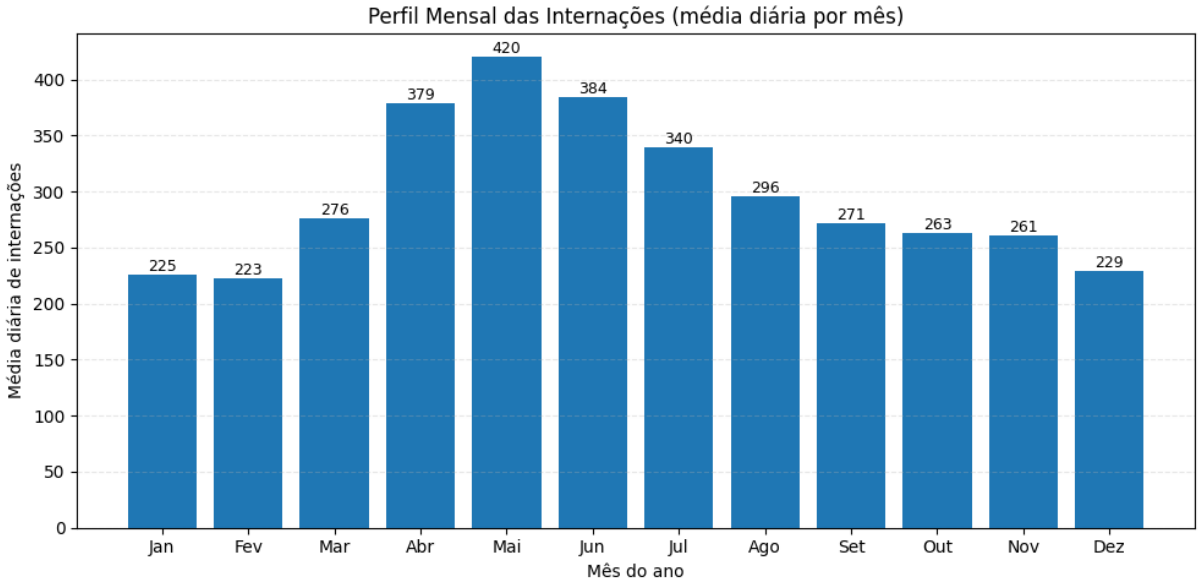

Figura 15: Perfil mensal das internações, obtido pela média do número de internações diárias em cada mês. Observa-se sazonalidade anual marcada, com aumento a partir de março, pico no outono/inverno (máximo em maio) e queda gradual a partir de agosto, atingindo menores médias na primavera e início do verão.

Além do padrão sazonal anual, também foi analisada a variação das internações ao longo dos dias da semana. Para isso, a série diária foi agregada por dia da semana, calculando-se a média diária de internações em cada dia (segunda a domingo).

A Figura 16 apresenta o perfil semanal. Observa-se que as principais médias diárias ocorrem nos dias úteis, com destaque para a segunda-feira (cerca de 337 internações/dia), seguida por quarta e terça-feira. A partir de quinta e sexta-feira há uma leve redução, nos fins de semana a queda se torna mais evidente, quando as médias de sábado e domingo giram em torno de 249 e 235 internações/dia, respectivamente. Esse comportamento indica um padrão semanal marcado, com maior carga de internações concentrada nos dias úteis e redução significativa aos finais de semana.

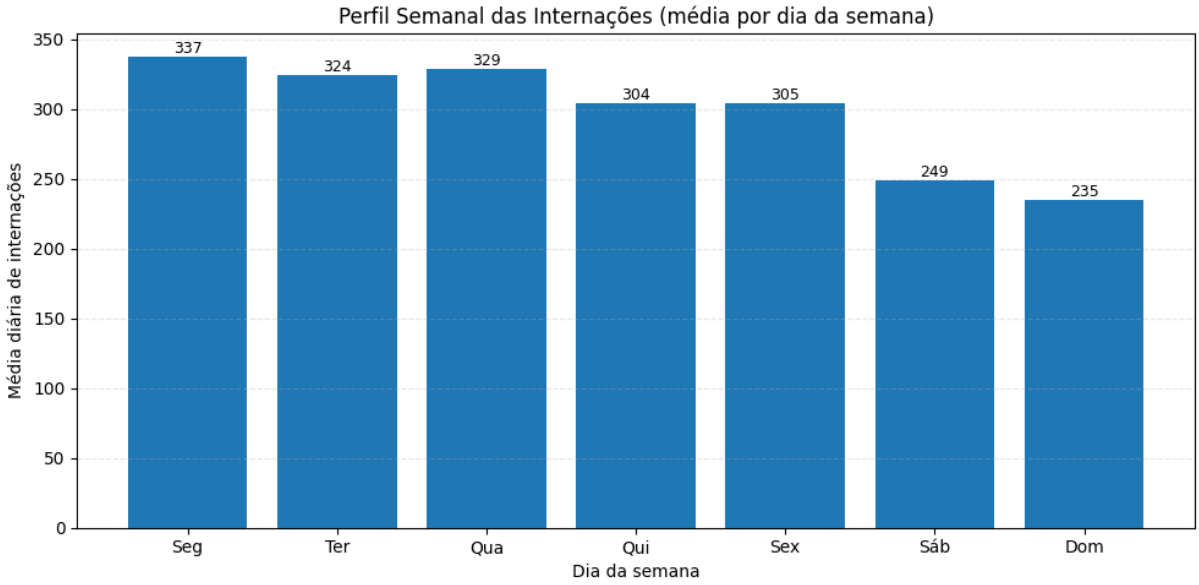

Figura 16: Perfil semanal das internações, obtido pela média do número de internações diárias em cada dia da semana. Observa-se maior média nos dias úteis, com pico na segunda-feira, e redução progressiva em direção ao fim de semana, com as menores médias no sábado e no domingo.

Por fim, investigou-se a tendência de longo prazo da série por meio da decomposição STL (Figura 17). Esse procedimento permite decompor a série observada em três partes principais: tendência de longo prazo, componente sazonal e componente residual.

No primeiro painel da Figura 17 é apresentada a série original de internações. Observa-se a combinação de grande variabilidade diária com picos recorrentes ao longo dos anos, além de uma percepção visual de queda gradual no nível médio da série.

O segundo painel mostra o componente de tendência estimado. Ele evidencia uma trajetória decrescente ao longo do período de estudo, com valores mais altos no início da série e redução progressiva nos anos seguintes. Essa tendência de longo prazo indica que, em média, o número diário de internações respiratórias vem diminuindo ao longo do tempo.

O terceiro painel exibe o componente sazonal, que representa o padrão anual recorrente após a remoção da tendência. Nota-se a presença de ciclos bem definidos, com amplitudes maiores em determinados períodos do ano, em linha com os resultados da análise de média mensal, sugerindo maior concentração de internações nos meses de outono e inverno.

Por fim, o quarto painel apresenta o componente residual, obtido após a remoção conjunta da tendência e da sazonalidade. Esse termo oscila em torno de zero e concentra flutuações de curto prazo e variações não explicadas pelos demais componentes. Observa-se que a amplitude desses resíduos tende a diminuir nos anos mais recentes, indicando redução da variabilidade

não estruturada da série.

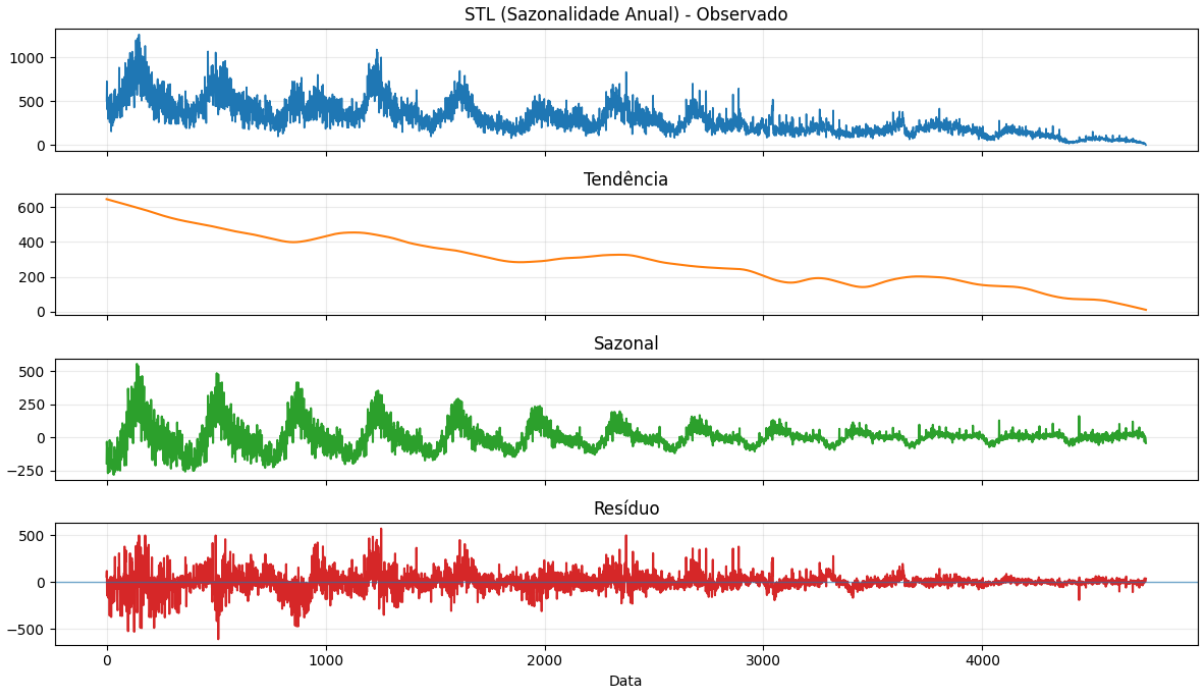

Figura 17: Decomposição STL aplicada à série diária de internações, separando o sinal observado em três componentes: tendência (variações de longo prazo no nível médio), sazonalidade anual (ciclo recorrente associado ao ano) e resíduo (flutuações de curto prazo não explicadas pelos demais componentes)

4.5.2 Dependência Temporal do Alvo

Além de padrões sazonais, séries temporais epidemiológicas frequentemente apresentam dependência temporal de curto prazo, na qual valores recentes influenciam observações futuras. Para investigar essa estrutura, utilizou-se a PACF, que permite identificar defasagens (*lags*) com contribuição temporal própria, controlando o efeito das defasagens intermediárias.

A PACF foi estimada utilizando a função `plot_pacf` da biblioteca *statsmodels*, considerando 30 defasagens e o método de Yule–Walker modificado (`method='ywm'`). A Figura 18 apresenta o resultado. Observa-se que as defasagens de **1 a 7 dias** apresentam autocorrelações parciais elevadas e estatisticamente significativas (fora do intervalo de confiança de 95%), indicando forte dependência do comportamento da última semana.

Adicionalmente, notam-se picos em *lags* múltiplos de 7 dias (por exemplo, 14, 21 e 28), reforçando a presença de um componente semanal na dinâmica do alvo. Esses achados motivam a inclusão de *features* autorregressivas (`lag_1` a `lag_7`) e de atributos relacionados à sazonali-

dade semanal na etapa de engenharia de atributos (Seção 4.7).

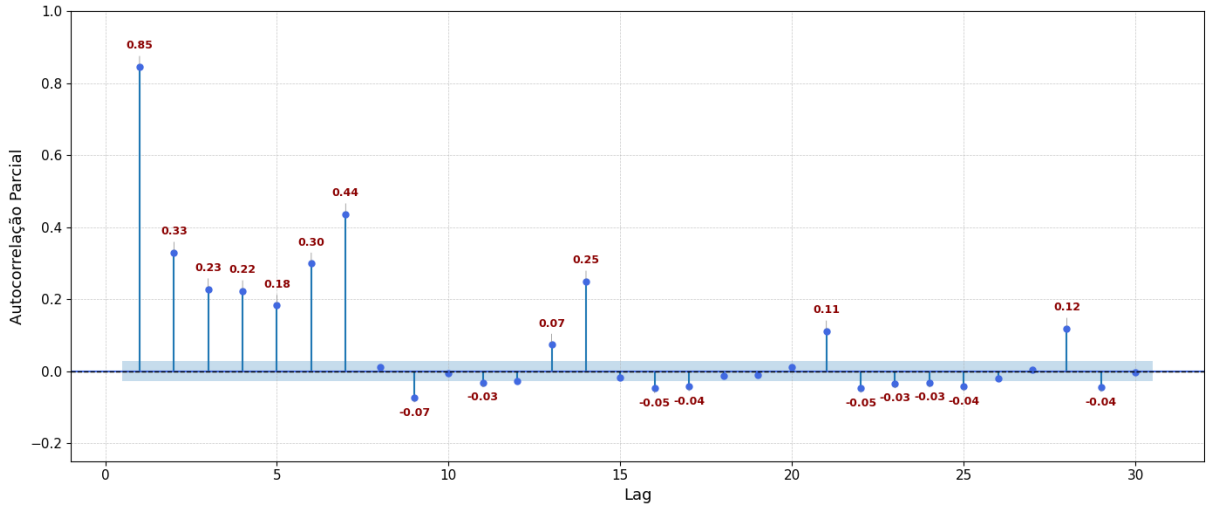

Figura 18: PACF da série diária de internações por doenças respiratórias, evidenciando autocorrelações significativas nas primeiras defasagens e em múltiplos de 7 dias.

Para avaliar o quanto as médias móveis das internações preservam a informação do número de hospitalizações, foi calculada a correlação de Pearson entre a série e diferentes janelas de média móvel ($3, 7, 14, 21, 30$ e 60 dias). Para cada janela, a média móvel foi calculada sobre a série `num_internacoes` e, em seguida, alinhada à série original, removendo-se os valores iniciais sem observação (valores `NaN`) gerados pelo cálculo da janela.

A Figura 19 apresenta os coeficientes de correlação obtidos. Observa-se que todas as médias móveis apresentam correlação elevada com a série original, com valores variando de aproximadamente 0,94 para a janela de 3 dias (`ma_3`) até 0,86 para a janela de 60 dias (`ma_60`).

À medida que a janela aumenta, a correlação diminui, refletindo que janelas mais longas suavizam e capturam tendências de longo prazo, enquanto janelas curtas preservam melhor as variações de curto prazo. Esses resultados indicam que médias móveis são candidatas relevantes a preditores, mas também sugerem a necessidade de selecionar um subconjunto de janelas para evitar redundância excessiva entre variáveis altamente correlacionadas.

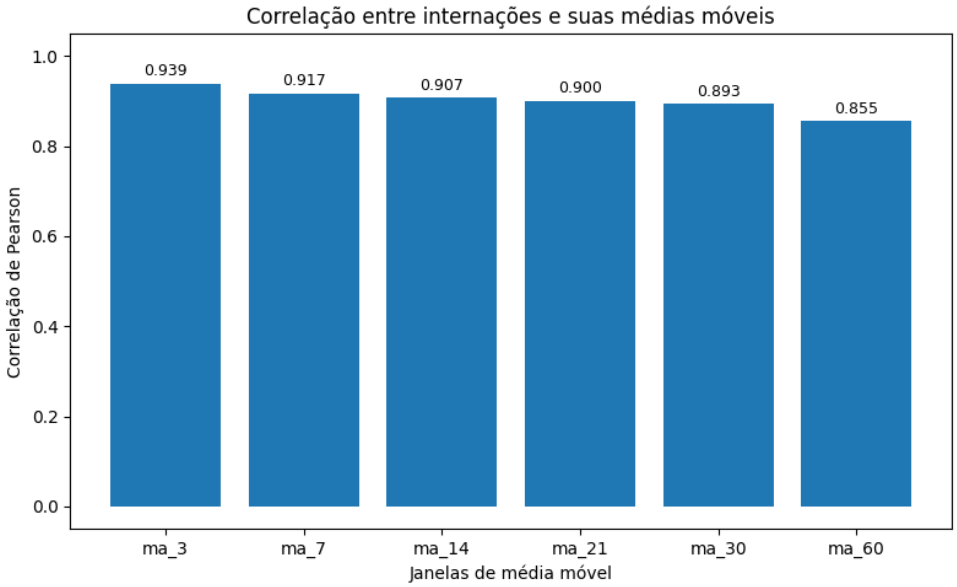

Figura 19: Correlação de Pearson entre a série diária de internações por doenças respiratórias e suas médias móveis com diferentes janelas (3, 7, 14, 21, 30 e 60 dias).

4.5.3 Análise das Variáveis Ambientais

Além da caracterização temporal do alvo, foi realizada uma análise exploratória das variáveis ambientais (meteorologia e poluentes) visando compreender suas distribuições, identificar assimetria e avaliar a presença de valores extremos. Essa etapa é importante porque séries ambientais, sobretudo de poluentes, frequentemente apresentam caudas longas e *outliers*, o que pode influenciar tanto medidas estatísticas (por exemplo, correlações) quanto o desempenho e a estabilidade de modelos preditivos.

De modo geral, observou-se que variáveis meteorológicas como `temp` e `ur` tendem a apresentar maior concentração em torno de valores centrais e menor assimetria, enquanto variáveis associadas à qualidade do ar exibem maior dispersão e ocorrência de extremos. A Figura 20 sintetiza esse comportamento por meio de *boxplots* das variáveis meteorológicas e dos poluentes antes da etapa de transformação estatística. Nota-se que diversas séries de poluentes apresentam caudas longas e muitos pontos fora dos limites do *boxplot*, evidenciando assimetria pronunciada e a presença de valores extremos.

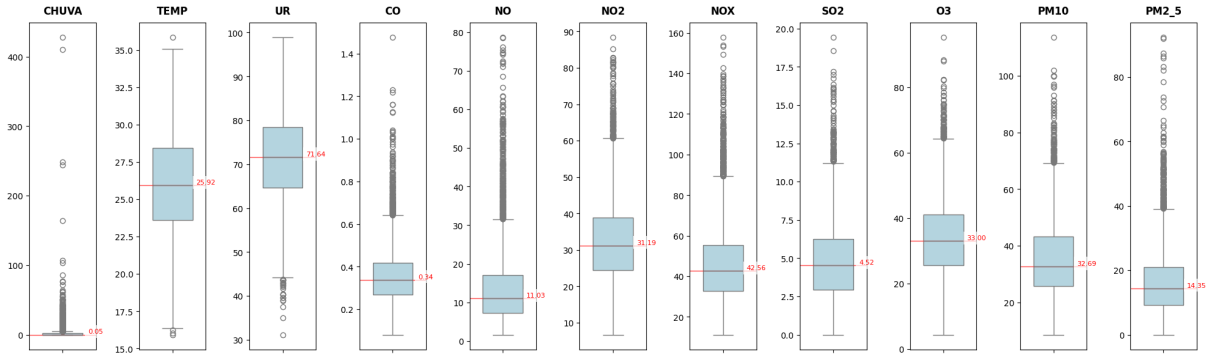

Figura 20: *Boxplots* das variáveis meteorológicas (`chuva`, `temp`, `ur`) e dos poluentes (`co`, `no`, `no2`, `nox`, `so2`, `o3`, `pm10`, `pm2_5`), evidenciando a presença de assimetria e valores extremos, principalmente nas séries de poluentes.

A análise dos *boxplots* motivou a aplicação de transformações de potência (Seção 4.4.3) com o intuito de reduzir assimetria, estabilizar a variância e mitigar a influência de *outliers*. Após a aplicação das transformações, as distribuições das variáveis foram reavaliadas. A Figura 21 apresenta os *boxplots* das variáveis transformadas, evidenciando redução da assimetria e menor frequência de extremos pronunciados. Em termos práticos, observa-se maior concentração dos valores na região central e diminuição da influência relativa de observações extremas, indicando maior estabilidade estatística das séries e tornando-as mais adequadas para análises exploratórias e para a etapa de modelagem.

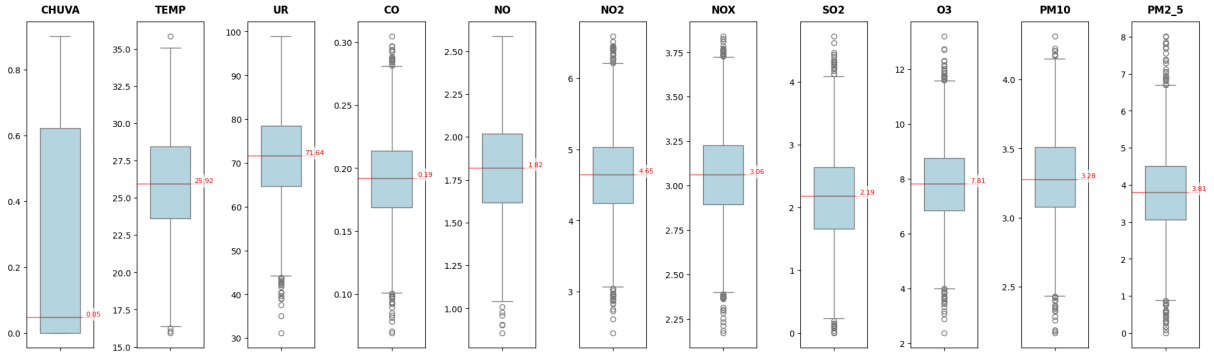

Figura 21: *Boxplots* das variáveis meteorológicas e poluentes após a etapa de transformação, mostrando redução da assimetria e da influência de valores extremos, especialmente nas séries de poluentes.

Em resumo, a análise exploratória das variáveis ambientais indicou que o tratamento estatístico aplicado contribuiu para tornar as distribuições mais regulares, reduzindo a assimetria e a presença de extremos.

4.6 Análise das Relações entre Variáveis Ambientais e Internações

Esta seção investiga associações entre as variáveis ambientais e o número diário de internações, utilizando medidas de correlação em diferentes estruturas temporais. O objetivo é identificar padrões de relacionamento (instantâneos, defasados e acumulados) que motivem a seleção de *features* ambientais e suas transformações para a etapa de modelagem (Seção 4.7). Como se trata de uma análise bivariada, os resultados devem ser interpretados como evidências de associação, não implicando causalidade.

4.6.1 Correlação Instantânea (Lag 0)

Visando medir a associação entre condições ambientais e internações, foi calculada a correlação de Spearman (ρ) entre a série diária de internações e as variáveis meteorológicas/poluentes no mesmo dia (lag 0). A escolha de Spearman se justifica por ser uma medida não-paramétrica, robusta a assimetria e a relações não necessariamente lineares, comportamento comum em séries ambientais.

A Figura 22 apresenta os coeficientes obtidos. Observa-se que os principais valores **positivos** ocorreram para os poluentes `no` ($\rho = 0{,}51$), `nox` ($\rho = 0{,}49$) e `no2` ($\rho = 0{,}41$), indicando associação moderada entre maiores concentrações desses óxidos de nitrogênio e maiores contagens de internações no mesmo dia. Em menor magnitude, `pm10` ($\rho = 0{,}25$) e `pm2_5` ($\rho = 0{,}18$) sugerem associação positiva mais fraca.

Entre as correlações **negativas**, destaca-se `o3` ($\rho = -0{,}36$), seguido de `so2` ($\rho = -0{,}22$) e `temp` ($\rho = -0{,}17$). A associação negativa com temperatura é coerente com a sazonalidade previamente observada, em que maiores médias de internações concentram-se em períodos mais frios do ano. Por fim, `chuva` ($\rho = -0{,}11$), `ur` ($\rho = -0{,}07$) e `co` ($\rho = 0{,}08$) exibiram coeficientes próximos de zero, sugerindo baixa associação instantânea no recorte bivariado.

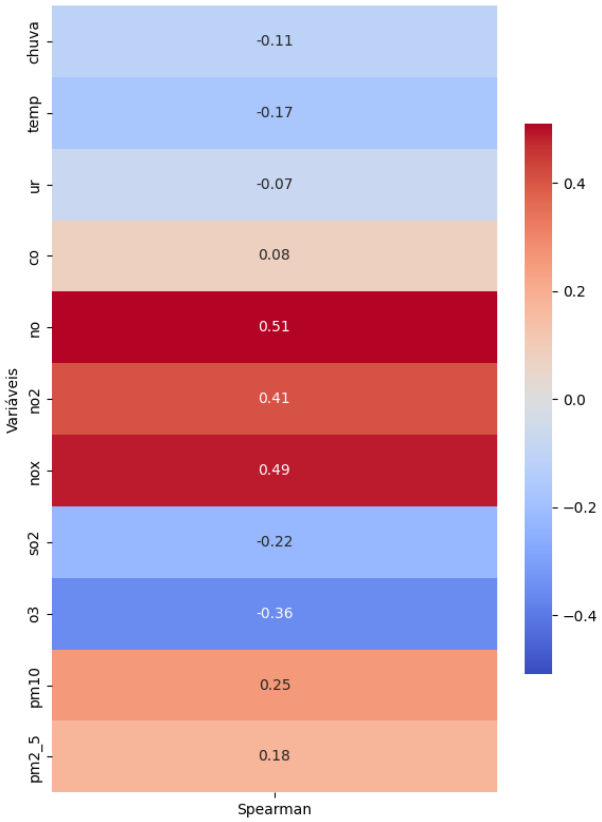

Figura 22: Mapa de calor dos coeficientes de correlação de Spearman (ρ) entre as variáveis meteorológicas e poluentes atmosféricos e o número diário de internações por doenças respiratórias, considerando correlação instantânea (lag 0). Tons de vermelho indicam associação positiva e tons de azul indicam associação negativa; os valores exibidos em cada célula correspondem ao ρ estimado para cada variável.

4.6.2 Correlações Defasadas (*lags* ambientais)

A análise de correlação instantânea (*lag* 0) fornece apenas uma visão do dia entre poluentes e internações. No entanto, é plausível que a resposta em saúde ocorra com algum atraso em relação à exposição ambiental (por exemplo, devido ao tempo de agravamento dos sintomas, busca por atendimento e efetiva internação). Para investigar esse comportamento, foram avaliadas correlações defasadas entre as variáveis ambientais e o alvo, deslocando-se cada poluente em k dias e correlacionando $X(t-k)$ com $Y(t)$.

Nesta etapa, foram considerados apenas os poluentes que apresentaram associação relevante na etapa anterior, sendo testadas defasagens $k \in \{1, 7, 12\}$ dias. Para cada variável X, foi construída a versão defasada X_{t-k}, alinhando-a à série de internações Y_t. Em seguida, calculou-se a correlação de Spearman (ρ) entre X_{t-k} e Y_t.

A Figura 23 resume os resultados em forma de *heatmap*. Observa-se um padrão consistente

de associação positiva para os **óxidos de nitrogênio**, com destaque para `no`, que apresentou o maior coeficiente em $k = 7$ ($\rho = 0{,}45$), seguido por `nox` ($\rho = 0{,}37$ em $k = 7$) e `no2` ($\rho = 0{,}28$ em $k = 7$). Em geral, nota-se que, para essas variáveis, a correlação é **mais alta em torno de uma semana** e permanece positiva também em $k = 1$ e $k = 12$, sugerindo que a dinâmica das internações pode estar associada a exposições recentes e/ou a condições ambientais persistentes ao longo de dias consecutivos.

Em contraste, `o3` manteve correlação negativa em todas as defasagens, com maior magnitude novamente em $k = 7$ ($\rho = -0{,}37$). Para `so2`, também se observou correlação negativa, com leve aumento em módulo à medida que o atraso cresce (de $-0{,}24$ em $k = 1$ para $-0{,}28$ em $k = 12$). Por fim, `pm10` e `pm2_5` apresentaram correlações positivas mais fracas (em torno de 0,12 a 0,16), sem variação expressiva entre os *lags* analisados.

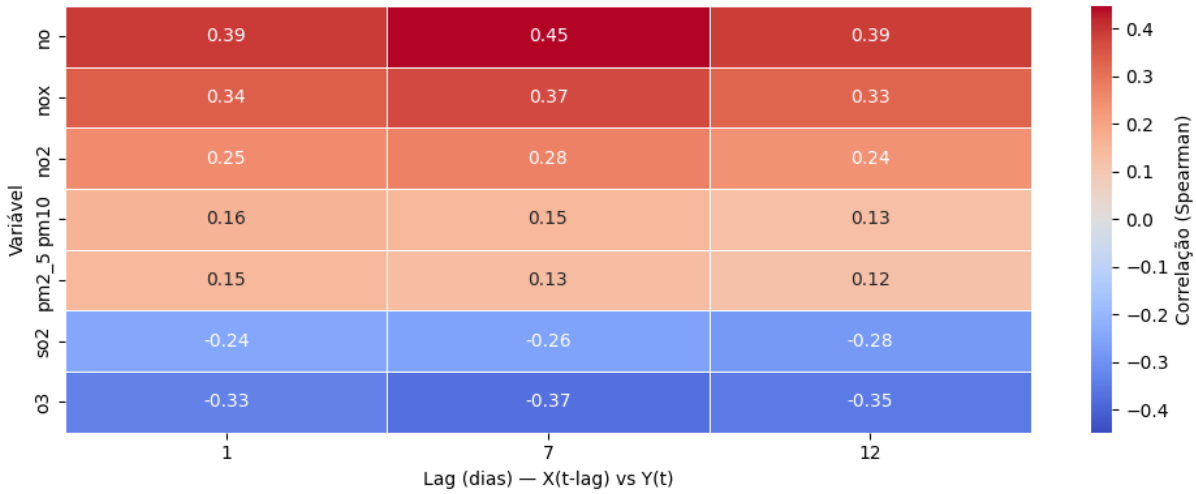

Figura 23: Mapa de calor dos coeficientes de Spearman entre poluentes defasados $X(t-k)$ e internações $Y(t)$, para $k \in \{1, 7, 12\}$ dias; valores anotados indicam ρ para cada poluente e defasagem.

4.6.3 Médias Móveis Defasadas e Exposição Acumulada

Como a resposta em saúde pode refletir não apenas a exposição pontual, mas também efeitos acumulados e retardados, foi avaliada uma abordagem baseada em médias móveis defasadas das variáveis ambientais. Para cada variável atmosférica, foram geradas médias móveis com janelas $w \in \{3, 7, 14, 21, 30, 60, 90, 120, 150\}$ dias. Para evitar vazamento de informação, todas as médias móveis foram calculadas de forma causal, com deslocamento temporal de pelo menos 1 dia, garantindo que a janela não incluísse o dia da previsão.

Em seguida, foi calculada a correlação de Spearman entre cada média móvel defasada e a série de internações, buscando identificar combinações (w, s) (janela e deslocamento) que maximizassem $|\rho|$. A Figura 24 apresenta, para cada variável, o maior coeficiente observado sob essa estratégia.

Os resultados indicam que, para alguns poluentes, a suavização temporal está associada a correlações mais altas do que as medidas diárias brutas. Entre as associações **positivas**, destaca-se `no` com $\rho = 0{,}59$ ao considerar a média móvel de 60 dias (deslocamento de 1 dia), seguido por `nox` ($\rho = 0{,}45$, também com janela de 60 dias) e `no2` ($\rho = 0{,}33$, com janela de 30 dias). Esse padrão sugere que, para óxidos de nitrogênio, a exposição acumulada em escalas de aproximadamente 1–2 meses acompanha mais de perto a variação das internações do que a exposição pontual no dia.

Nas associações **negativas**, observa-se maior correlação em módulo para `o3`, com $\rho = -0{,}68$ em uma média móvel de 150 dias (deslocamento de 1 dia), e para `so2`, cuja maior correlação ocorreu em janela longa e com deslocamento adicional ($\rho = -0{,}53$). Em contraste, `co` permaneceu com correlação próxima de zero mesmo após a suavização. De forma geral, esses achados sugerem que diferentes variáveis podem se relacionar ao desfecho em escalas temporais distintas.

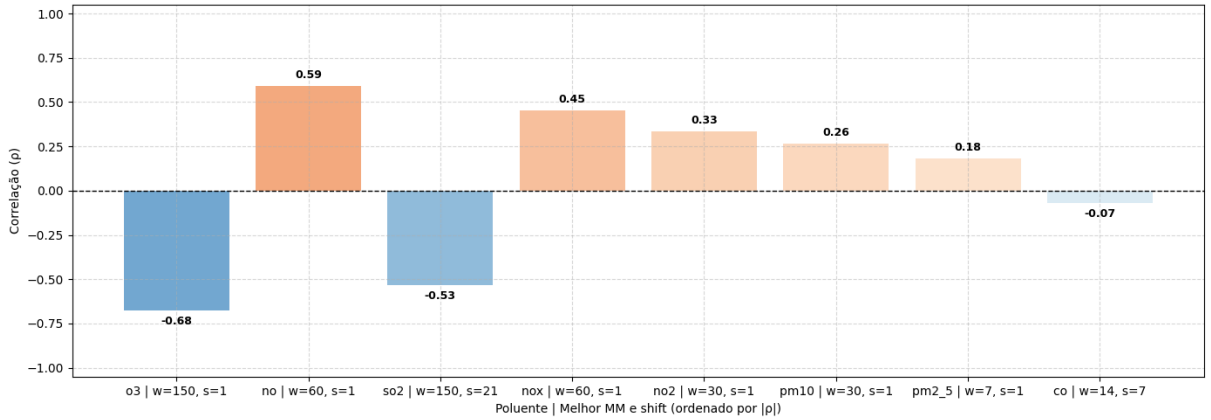

Figura 24: Maior correlação de Spearman (ρ) entre as internações e médias móveis defasadas das variáveis ambientais. Para cada variável, selecionou-se a combinação de janela (w) e deslocamento (s) que maximizou $|\rho|$.

4.6.4 Interpretação dos Padrões Observados

Em conjunto, as análises indicam que:

- Associações instantâneas (*lag* 0) são mais evidentes para óxidos de nitrogênio (`no`, `no2`,

nox), enquanto outras variáveis apresentam correlação fraca no mesmo dia.

- Exposição acumulada pode estar melhor representada por médias móveis em janelas de semanas a meses para certos poluentes, sugerindo que a dinâmica do desfecho pode responder a padrões ambientais persistentes, e não apenas a variações diárias.

- Correlação negativa observada para o3 (e, em menor grau, temp) pode refletir fatores compartilhados de sazonalidade e meteorologia: ozônio tende a se elevar em condições mais quentes e ensolaradas, enquanto internações respiratórias se concentram em períodos mais frios. Assim, a associação negativa não deve ser interpretada como efeito protetor, mas como uma possível *covariação* induzida por sazonalidade/tendência.

Por se tratarem de análises bivariadas, correlações podem incorporar simultaneamente efeitos diretos, efeitos de confusão (por sazonalidade, tendência e padrões semanais) e dependência temporal do próprio alvo. Por esse motivo, os padrões identificados nesta seção foram utilizados como critérios de priorização para a seleção e construção de *features* ambientais (defasagens e médias móveis) na Seção 4.7, em conjunto com controles sazonais e *features* autorregressivas de internações, visando reduzir o risco de interpretações espúrias e melhorar a capacidade preditiva do modelo.

4.7 Engenharia de Atributos

Esta seção descreve a engenharia de atributos realizada a partir do *dataset* diário unificado (Seção 4.4.4). O objetivo foi transformar a série-alvo (internações) e as variáveis ambientais em representações numéricas capazes de capturar: (i) dependência temporal de curto prazo, (ii) padrões sazonais (semanal e anual), (iii) sinais estruturados derivados da decomposição STL e (iv) possíveis efeitos defasados e acumulados de meteorologia e poluentes. Em todas as construções, adotou-se um regime **estritamente causal** (isto é, as *features* em t utilizam apenas informação disponível até $t-1$), evitando vazamento de informação.

4.7.1 Features Autorregressivas das Internações (*lags* + médias móveis)

Com base nas evidências de dependência temporal de curto prazo observadas na análise de autocorrelação parcial (Seção 4.5.2), foram criadas *features* autorregressivas a partir da própria

série diária de internações. Especificamente, foram consideradas defasagens de 1 a 7 dias:

$$\texttt{lag\_k}(t) = y_{t-k}, \quad k \in \{1, \ldots, 7\},$$

em que y_t representa o número de internações no dia t. Essas variáveis (`lag_1`, ..., `lag_7`) permitem que o modelo incorpore a dinâmica recente da série-alvo como sinal preditivo.

Além das defasagens pontuais, foram incluídas *features* de suavização de curto prazo por meio de médias móveis causais. Foram calculadas médias móveis com janelas $w \in \{3, 7\}$, utilizando deslocamento temporal (`shift(1)`) para impedir que o dia corrente participe do cálculo:

$$\texttt{ma\_w}(t) = \frac{1}{w} \sum_{i=1}^{w} y_{t-i}.$$

Assim, `ma_3` e `ma_7` representam a intensidade recente do nível de internações, reduzindo ruído diário e auxiliando o modelo a responder a alterações persistentes.

4.7.2 *Features* de Sazonalidade (mês, semana, codificação cíclica)

As análises de sazonalidade mensal e semanal (Seção 4.5.1) indicaram padrões recorrentes associados ao calendário. Para representar esses ciclos sem introduzir descontinuidades artificiais (por exemplo, entre dezembro e janeiro), foram criadas *features* cíclicas a partir do mês do ano e do dia da semana.

Seja $m(t) \in \{1, \ldots, 12\}$ o mês e $d(t) \in \{0, \ldots, 6\}$ o dia da semana. As codificações cíclicas foram definidas por:

$$\texttt{mes\_sin}(t) = \sin\left(2\pi \frac{m(t)}{12}\right), \quad \texttt{mes\_cos}(t) = \cos\left(2\pi \frac{m(t)}{12}\right),$$

$$\texttt{dow\_sin}(t) = \sin\left(2\pi \frac{d(t)}{7}\right), \quad \texttt{dow\_cos}(t) = \cos\left(2\pi \frac{d(t)}{7}\right).$$

Adicionalmente, foram criadas variáveis indicadoras para refletir padrões específicos observados nos dados: `is_weekend`, indicando sábados e domingos, e `is_monday`, destacando segundas-feiras, dia com média elevada de internações no período analisado.

4.7.3 *Features* da Decomposição STL (componente anual e amplitude)

Para capturar um sinal sazonal anual mais estruturado do que simples variáveis de calendário, utilizou-se a decomposição STL descrita na Seção 4.5.1. A motivação é que *features* de calendário indicam *posição* no ciclo (mês/dia da semana), mas não informam a *intensidade* local da sazonalidade ou eventuais mudanças no padrão sazonal ao longo do tempo.

A série diária de internações foi decomposta em janelas móveis (*rolling*) em regime causal, com período anual de 365 dias. Para cada dia t, a STL foi ajustada sobre uma janela histórica de aproximadamente três anos (cerca de 1095 dias), exigindo histórico mínimo para estimar uma sazonalidade anual estável. A partir dessa decomposição, foram extraídas duas *features* principais, sempre com deslocamento para garantir causalidade:

- `stl365_seasonal`: valor da componente sazonal anual no dia $t-1$, funcionando como um *baseline* sazonal esperado para o período.
- `stl365_season_amp`: amplitude sazonal local, definida como a diferença entre o máximo e o mínimo da componente sazonal nos últimos 365 dias (utilizando apenas informação até $t-1$), quantificando a "força" do ciclo anual no intervalo recente.

Essas variáveis sintetizam a sazonalidade anual em dois aspectos complementares: nível (componente sazonal) e intensidade (amplitude), permitindo que o modelo distinga períodos com sazonalidade mais pronunciada de períodos mais estáveis.

4.7.4 *Features* Meteorológicas e de Poluentes

Com base nas análises de associação entre variáveis ambientais e internações (Seção 4.6), foram definidas *features* para representar: (i) exposição contemporânea, (ii) efeitos defasados de curto prazo e (iii) exposição acumulada.

Inicialmente, foi realizada uma triagem pela correlação de Spearman no mesmo dia (lag 0) entre cada variável ambiental e o número de internações, adotando-se como critério $|\rho| \geq 0{,}30$. Com isso, foram selecionadas as variáveis `no`, `no2`, `nox` e `o3`, incorporadas como *features* contemporâneas (`no_t`, `no2_t`, `nox_t` e `o3_t`).

Para capturar efeitos combinados e potenciais não-linearidades na mistura de poluentes, foram incluídas *features* de interação quando as respectivas variáveis estavam disponíveis. Em

particular, criou-se `nox_x_o3_t` quando `nox` e `o3` estavam presentes, motivada pelo comportamento observado de sinais opostos e pela hipótese de que a combinação pode carregar informação adicional em relação às séries isoladas.

Para modelar a possibilidade de impactos tardios entre exposição e desfecho respiratório, foram criadas defasagens para variáveis ambientais selecionadas, contemplando horizontes curtos e semanais (por exemplo, $k = 1$ e $k = 7$):

$$\texttt{var\_lagk}(t) = x_{t-k}.$$

Foram incluídas, por exemplo, `no_lag1`, `no_lag7`, `o3_lag1` e `o3_lag7`, bem como defasagens específicas para outras variáveis (por exemplo, `pm10_lag1`, `pm2_5_lag1` e `so2_lag12`), conforme indicado pelas análises exploratórias.

Por fim, para representar o efeito acumulado de exposição ao longo de uma janela recente, foram consideradas médias móveis defasadas para variáveis selecionadas. Para uma variável ambiental x_t, utilizou-se a construção causal com deslocamento $s \geq 1$ antes do cálculo da média em janela w:

$$\texttt{ma\_x\_w\_s}(t) = \frac{1}{w} \sum_{i=s}^{s+w-1} x_{t-i}.$$

Essa parametrização permite testar diferentes combinações de janela (ex.: 3, 7, 14 dias) e deslocamento (ex.: 1, 2, 7 dias), garantindo que a *feature* em t não incorpore informação do próprio dia da previsão.

Em conjunto, as *features* descritas nesta seção compõem a matriz final de entrada do modelo, integrando sinais autorregressivos, sazonais e ambientais de forma consistente com o cenário de previsão. A próxima seção descreve o treinamento, validação e avaliação do modelo preditivo utilizando esse conjunto de atributos.

4.7.5 Construção Final da Matriz de Features

Após a criação dos atributos descritos, foi construída a matriz final de preditores X e o vetor-alvo y utilizados no treinamento e na avaliação do modelo. O alvo foi definido como a contagem diária de internações respiratórias, $y_t = \texttt{num\_internacoes}(t)$. Para cada dia t, a linha correspondente em X_t reúne: (i) *lags* e médias móveis das internações, (ii) variáveis de sazonalidade (calendário e codificação cíclica), (iii) atributos derivados da decomposição STL

e (iv) variáveis meteorológicas e de poluentes, incluindo versões defasadas e agregadas quando aplicável.

A combinação de *lags*, médias móveis e principalmente dos atributos STL (`stl365_seasonal` e `stl365_season_amp`) implica a geração de valores ausentes (`NaN`) no início da série, decorrentes da necessidade de histórico mínimo para o cálculo dessas variáveis. Na prática, observou-se que a principal perda de observações foi causada pelas *features* STL, uma vez que sua estimação exige uma janela extensa de dados passados para estabilizar a componente sazonal anual. Assim, para construir um conjunto consistente, foram removidas as linhas iniciais do período em que essas *features* não estavam disponíveis.

Como resultado, embora o *dataset* diário unificado cobrisse originalmente o intervalo de 2012 a 2024, a **matriz final de *features*** utilizada na modelagem passou a contemplar o período de **2014 a 2024**.

4.8 Desenvolvimento do Modelo Preditivo

A partir do *dataset* final construído nas etapas de pré-processamento e engenharia de atributos (Seções 4.4 e 4.7), foi desenvolvido um modelo preditivo para estimar o número diário de internações por doenças respiratórias. O problema foi tratado como uma tarefa de regressão supervisionada em série temporal, na qual a matriz de preditores combina sinais autorregressivos (*lags* e médias móveis do alvo), atributos sazonais (calendário e STL) e variáveis meteorológicas e de poluentes (valores contemporâneos, defasagens e exposição acumulada). Nesta seção descreve-se a formulação do problema, a estratégia de particionamento temporal, o *detrending* causal do alvo, a seleção de variáveis, o ajuste do modelo e os critérios de avaliação.

4.8.1 Estratégia de Treino/Teste e Validação Temporal

Como se trata de um problema de série temporal, todo o processo de treinamento e avaliação respeitou a ordem cronológica, evitando embaralhamento aleatório de amostras. A base foi organizada em frequência diária contínua, e o conjunto de teste foi definido como o trecho mais recente da série, seguindo um particionamento *hold-out* temporal.

Adicionalmente, foi aplicado um *gap* (em dias) entre os conjuntos de treino e teste para reduzir dependências de curto prazo e possíveis contaminações induzidas por *features* construídas via janelas móveis e defasagens. Para o ajuste de hiperparâmetros e seleção de variáveis, foi

empregada validação cruzada temporal com múltiplos *folds* ordenados no tempo (por exemplo, janelas expansivas ou deslizantes), de modo que cada validação simulasse um cenário de previsão em que o modelo é treinado no passado e avaliado em um bloco futuro.

4.8.2 *Detrending* Causal do Alvo

As análises exploratórias indicaram a presença de tendência de longo prazo na série de internações (Seção 4.5.1). Para reduzir o impacto de mudanças lentas e melhorar a capacidade do modelo de capturar variações de curto e médio prazo, foi adotada uma estratégia de *detrending* causal do alvo.

O procedimento consiste em estimar uma tendência τ_t por meio de uma média móvel não centrada (isto é, calculada apenas com observações anteriores) e, em seguida, treinar o modelo para prever o residual:

$$y_t^{\text{res}} = y_t - \tau_t.$$

Na etapa de previsão, o valor final é recomposto somando-se a tendência estimada:

$$\hat{y}_t = \widehat{y_t^{\text{res}}} + \tau_t.$$

Como τ_t é calculada utilizando exclusivamente informações disponíveis até $t-1$, preserva-se a aderência ao cenário real de previsão e evita-se vazamento de informação. Experimentos preliminares indicaram melhora relevante de desempenho quando o RF foi treinado sobre o alvo detrendido, razão pela qual essa estratégia foi incorporada ao fluxo metodológico.

4.8.3 Seleção de Variáveis

A combinação de *lags*, médias móveis, atributos sazonais e variáveis ambientais pode resultar em um conjunto de preditores com dimensionalidade elevada e redundâncias. Para reduzir ruído e favorecer generalização, foi aplicada uma etapa de seleção automática de variáveis.

A estratégia adotada foi baseada em importância por permutação(*permutation importance*) calculada em validações temporais. Em linhas gerais, mede-se o quanto o erro de previsão se degrada ao embaralhar uma variável candidata, preservando as demais. Para privilegiar o regime mais próximo do período de teste, foi dada maior ênfase aos *folds* mais recentes. Ao final, manteve-se apenas um subconjunto com as k variáveis mais relevantes, que alimentou o

treinamento e o ajuste do modelo.

4.8.4 Treinamento, Ajuste de Hiperparâmetros e Previsão

O treinamento do RF foi realizado para equilibrar capacidade preditiva e generalização. O ajuste de hiperparâmetros foi conduzido por busca aleatória (*Randomized Search*), contemplando parâmetros clássicos do método, como: número de árvores, profundidade máxima, número mínimo de amostras por folha e critérios de divisão. O critério de seleção foi o MAE, por ser robusto e diretamente interpretável na unidade do problema (internações/dia).

Para aumentar a aderência ao comportamento recente da série (mais próximo do período de teste), adotou-se um esquema de ponderação por recência durante o ajuste, atribuindo pesos maiores a observações mais recentes. Essa decisão é útil em contextos com mudanças graduais ao longo dos anos, pois induz o modelo a calibrar-se melhor no regime temporal mais atual.

4.8.5 Métricas de Avaliação e Procedimentos Diagnósticos

O desempenho do modelo foi avaliado separadamente em treino e teste, combinando métricas numéricas e diagnósticos gráficos (*Real vs. Previsto*). As métricas utilizadas foram:

- MAE e RMSE, para quantificar erro médio absoluto e penalização maior de erros grandes;
- R^2, para medir a proporção da variância do alvo explicada pelo modelo;
- sMAPE e wMAPE, para avaliar erro relativo de forma mais estável em diferentes magnitudes do alvo;
- **Bias** ($\mathbb{E}[\hat{y} - y]$), para identificar tendência sistemática de superestimação ou subestimação.

Como procedimento diagnóstico adicional, o viés foi analisado ao longo do calendário (por exemplo, por mês do ano), permitindo verificar se o modelo apresenta erros sistemáticos em períodos específicos, como meses de pico sazonal. Esse tipo de avaliação é particularmente relevante para o uso operacional do modelo em planejamento de demanda hospitalar, pois erros sazonais podem comprometer decisões em períodos críticos mesmo quando métricas agregadas são satisfatórias.

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos com o modelo preditivo proposto para estimar o número diário de internações por doenças respiratórias. Os resultados são analisados quantitativamente, com base em métricas de erro no conjunto de teste, e qualitativamente, por meio de gráficos *Real vs. Previsto*. Além disso, são apresentados resultados de interpretabilidade via SHAP, visando identificar quais grupos de variáveis mais contribuem para as previsões e verificar se o modelo captura padrões temporais relevantes.

5.1 Plano de Experimentos

Esta seção descreve como os experimentos foram organizados e estruturados, funcionando como uma ponte entre a metodologia e os resultados. Em particular, detalha-se (i) o que foi testado, (ii) por que foi testado, (iii) como os testes foram configurados no tempo e (iv) quais comparações serão realizadas nas subseções seguintes.

5.1.1 Objetivo dos experimentos

A etapa experimental teve como objetivos principais:

- Avaliar o desempenho preditivo do modelo proposto (RF) no horizonte de previsão de 1 dia.
- Comparar estratégias de modelagem, especialmente o treinamento direto sobre o alvo original em relação ao o treinamento com *detrending* causal do alvo, verificando se a remoção de tendência melhora a generalização em períodos recentes.
- Verificar a contribuição das *features* (autorregressivas, sazonais e ambientais), tanto via desempenho preditivo quanto via interpretabilidade (SHAP).
- Analisar se o modelo captura padrões sazonais e variações de curto prazo, utilizando diagnósticos visuais *Real vs. Previsto* para identificar aderência a oscilações e mudanças

de patamar.

5.1.2 Descrição dos cenários testados

Foram avaliados dois cenários experimentais, mantendo-se o mesmo modelo e a mesma formulação de previsão diária. A diferença entre os cenários está na forma como o alvo é apresentado ao modelo durante o treinamento.

Cenário 1 — Modelo sem `detrend` (*baseline*). Neste cenário, o modelo é treinado diretamente para prever o alvo original $y_t = \texttt{num\_internacoes}(t)$, sem remoção explícita de tendência. O conjunto de preditores inclui as *features* definidas na metodologia: *lags* e médias móveis das internações, variáveis sazonais (calendário e componentes STL), bem como variáveis ambientais (valores diários, versões defasadas e médias móveis defasadas). Esse cenário é incluído como *baseline* por representar a abordagem mais direta e por permitir avaliar o quanto a presença de tendência pode prejudicar a capacidade do modelo de generalizar para o período mais recente.

Cenário 2 — Modelo com `detrend` causal do alvo. Neste cenário, aplica-se o *detrending* causal descrito na Seção 4.8.2. No geral, estima-se uma tendência de longo prazo τ_t utilizando apenas informações passadas e treina-se o modelo para prever o residual $y_t^{\text{res}} = y_t - \tau_t$. Na fase de teste, a previsão final é recomposta como $\hat{y}_t = \widehat{y_t^{\text{res}}} + \tau_t$, retornando a escala original de internações por dia. A expectativa neste cenário é reduzir o impacto de mudanças lentas de regime e favorecer a aprendizagem de padrões sazonais e variações de curto prazo, levando a melhor calibração e menor viés no período de teste.

5.1.3 Conjunto de treino, teste e configuração temporal

Todos os experimentos respeitaram a ordem cronológica, sem embaralhamento de amostras, com horizonte de previsão de 1 dia. O particionamento foi feito no formato *hold-out* temporal, com o conjunto de teste definido como o trecho mais recente da série e o conjunto de treino composto pelos dados anteriores. Além disso, foi aplicado um *gap* temporal entre treino e teste para reduzir dependências de curto prazo e eventuais contaminações decorrentes de janelas móveis e defasagens utilizadas na construção das *features*. O procedimento de validação temporal empregado durante ajuste e seleção de variáveis segue o descrito na Seção 4.8.1.

5.1.4 Análises complementares

Além das métricas globais no conjunto de teste, os resultados serão complementados por:

- **Gráficos *Real vs. Previsto***, para diagnóstico qualitativo, avaliação de calibração de nível, captura de tendência/sazonalidade e resposta a oscilações de curto prazo.
- **Análise de interpretabilidade via SHAP**, para identificar quais atributos (autorregressivos, sazonais e ambientais) exercem maior influência nas previsões e verificar se os padrões apontados nas análises exploratórias e de correlação se refletem na estrutura explicativa do modelo.

As subseções seguintes apresentam os resultados de cada cenário, com comparações diretas entre desempenho, comportamento temporal das previsões e importância das variáveis.

5.2 Resultados do modelo sem `detrend`

Nesta seção são apresentados os resultados do RF treinado diretamente sobre o alvo original (sem remoção explícita de tendência). A Tabela 9 resume as métricas no conjunto de teste.

Tabela 9: Métricas de desempenho do RF no conjunto de teste ao prever diretamente o alvo original (`num_internacoes`), sem remoção de tendência (*detrending*).

Métrica	**Valor**
MAE (internações/dia)	31.03
RMSE (internações/dia)	39.36
R^2	0.37
sMAPE (%)	38.23
wMAPE (%)	33.04
Bias ($\hat{y} - y$, internações/dia)	24.81

De forma geral, observa-se um desempenho limitado no período de teste ($R^2 = 0{,}37$), com erros absolutos elevados (MAE de 31 internações/dia e RMSE de 39 internações/dia). Além disso, o *Bias* positivo (24,81) indica uma tendência sistemática do modelo a superestimar as internações no teste, ou seja, produzir previsões acima dos valores observados.

A Figura 25 apresenta a comparação *Real x Previsto* no conjunto de teste. Nota-se que, em parte do início do período, o modelo acompanha algumas oscilações de curto prazo, mas

no trecho final a previsão torna-se praticamente quase constante, mantendo-se em torno de um patamar aproximado enquanto a série observada apresenta queda e variações.

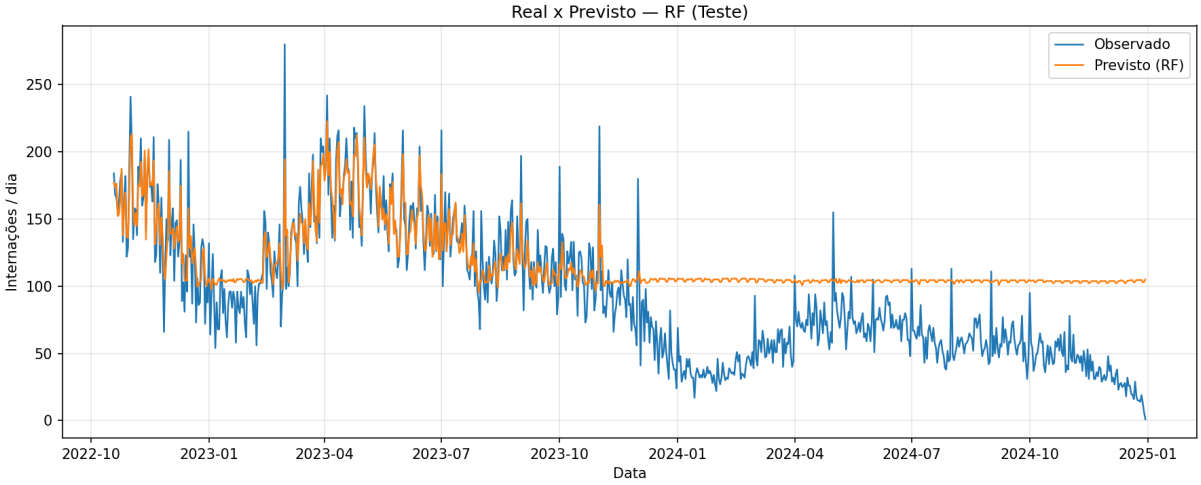

Figura 25: Comparação entre a série observada e as previsões do RF no conjunto de teste, sem aplicação de *detrending* do alvo.

A Figura 26 mostra o resumo dos valores SHAP do modelo no cenário sem `detrend`. Observa-se que o modelo depende principalmente de *features* autorregressivas da própria série de internações, o que é coerente com a forte dependência temporal de curto prazo identificada anteriormente (PACF e médias móveis). Em termos interpretativos, valores mais altos dessas variáveis tendem a empurrar a previsão para cima (SHAP positivo), enquanto valores mais baixos tendem a reduzir a previsão (SHAP negativo), refletindo a lógica de persistência do processo.

Também aparecem sinais sazonais/semanais entre as variáveis mais influentes, enquanto *features* ambientais surgem com menor contribuição relativa neste ajuste. Esse padrão sugere que, sem tratar a tendência, o modelo prioriza fortemente o histórico recente das internações e pode ter dificuldade para se adaptar quando ocorre mudança de nível/regime no período de teste, motivando a análise do cenário com `detrend` na próxima seção.

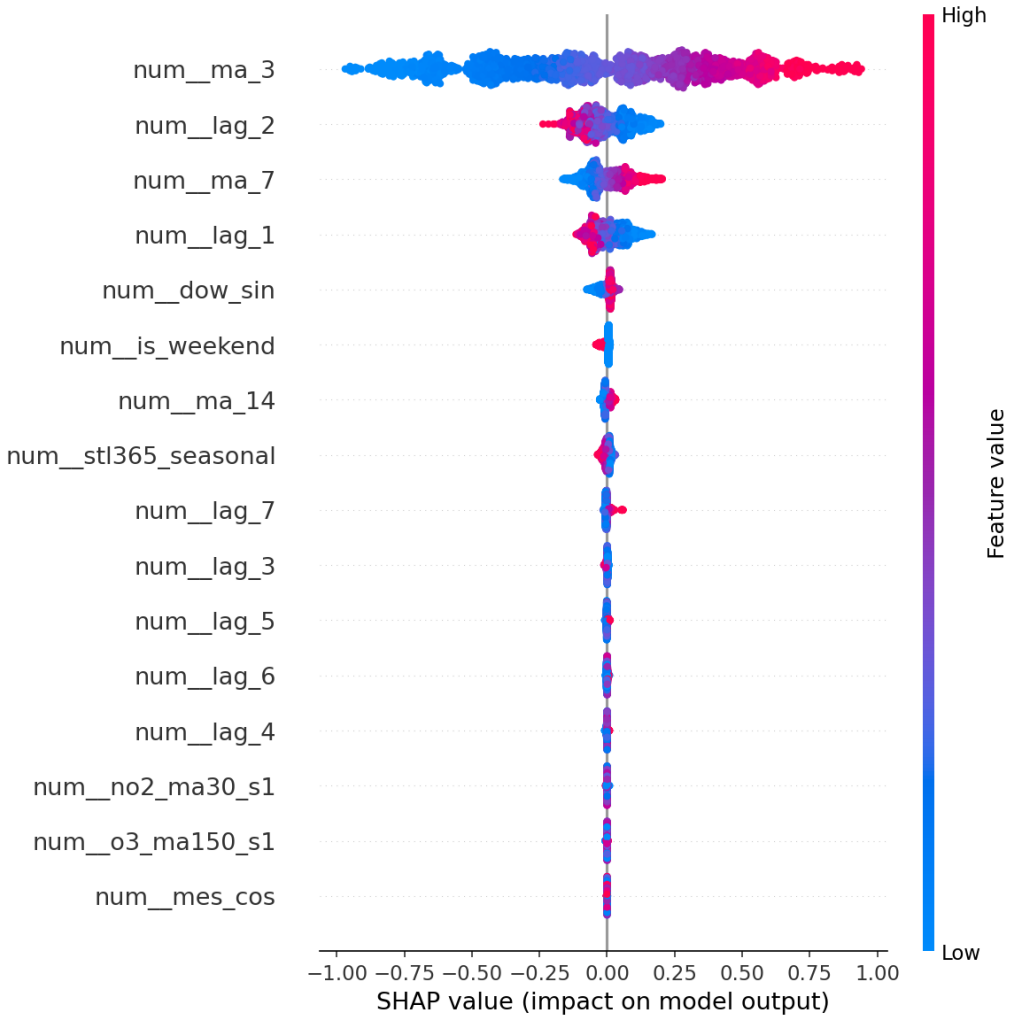

Figura 26: Resumo dos valores SHAP do RF no cenário sem *detrending*. Cada ponto representa uma observação; a cor indica o valor da *feature* (baixo/alto) e a posição no eixo *x* representa o impacto na saída do modelo.

5.3 Resultados do modelo com `detrend`

Nesta seção são apresentados os resultados do RF treinado com *detrending* causal do alvo. Como discutido na Seção 4.8.2, essa estratégia busca reduzir a influência de tendências lentas na série de internações, permitindo que o modelo aprenda principalmente as variações de curto/médio prazo e efeitos sazonais. Ao final da previsão, a tendência estimada é recomposta, retornando as estimativas para a escala original de internações por dia.

A Tabela 10 resume bem o desempenho no conjunto de teste e, em comparação direta com o cenário sem `detrend` (Tabela 9), observa-se uma melhora consistente em todas as métricas. O MAE reduz de 31,03 para 22,55 (queda de aproximadamente 27%), e o RMSE reduz de 39,36 para 28,73 (queda de aproximadamente 27%). Além disso, o coeficiente de determinação au-

menta de $R^2 = 0{,}37$ para $R^2 = 0{,}663$, indicando ganho expressivo de explicação da variabilidade do sinal no período de teste. Outrossim, o viés médio também melhora substancialmente: sai de um *Bias* fortemente positivo (24,81, indicando superestimação sistemática) para um valor próximo de zero e levemente negativo (-5,29), sugerindo predições muito mais bem calibradas no nível médio.

Tabela 10: Métricas de desempenho do RF no conjunto de teste ao prever o alvo com *detrending* causal (modelo treinado sobre a série detrendida e previsão final obtida por recomposição da tendência).

Métrica	**Valor (Teste)**
MAE (internações/dia)	22,55
RMSE (internações/dia)	28,73
R^2	0,663
sMAPE (%)	30,58
wMAPE (%)	24,02
Bias ($\hat{y} - y$, internações/dia)	-5,29

A Figura 27 apresenta a comparação *Real vs. Previsto* no período de teste. Em comparação com o cenário sem `detrend` (Figura 25), nota-se que a curva prevista passa a acompanhar melhor a dinâmica temporal do sinal observado, capturando mudanças de patamar e oscilações de curto prazo com maior fidelidade. Porém, picos muito abruptos tendem a ser suavizados.

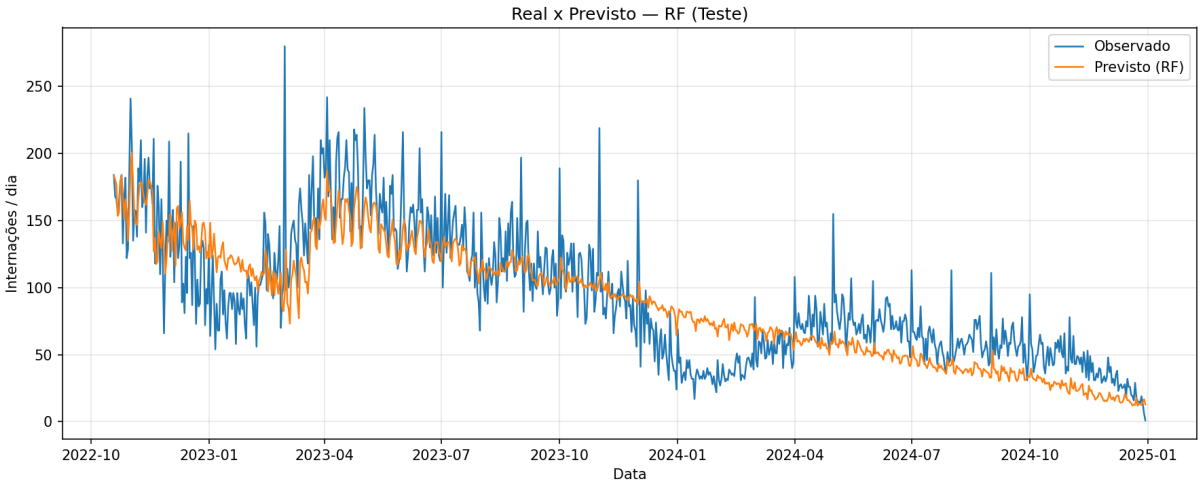

Figura 27: Comparação entre internações observadas e previstas pelo RF no conjunto de teste, utilizando *detrending* causal do alvo e posterior recomposição da tendência.

A Figura 28 apresenta o *summary plot* de SHAP do modelo com `detrend`. Ao comparar com o cenário sem *detrending* (Figura 26), observa-se que o modelo permanece dependendo por

sinais autorregressivos, porém atribui maior relevância a variáveis sazonais de maior abstração, como os termos derivados do STL (`stl365_seasonal` e `stl365_season_amp`). Isso pode sugerir que, ao remover a tendência, o modelo consegue explorar de forma mais eficiente a posição da série no ciclo anual e a intensidade do padrão sazonal.

Além disso, aparecem as variáveis associados a exposição acumulada/defasada (por exemplo, `so2_ma150_s21`, `pm10_ma30_s1`) e defasagens específicas, coerentes com as análises de correlação por médias móveis e lags (Seção 4.6).

Em resumo, esse padrão reforça que o `detrend` facilita a separação entre componentes de longo prazo e variação de curto/médio prazo, permitindo que o modelo incorpore de forma mais clara os sinais sazonais e ambientais sem degradar a generalização no período de teste.

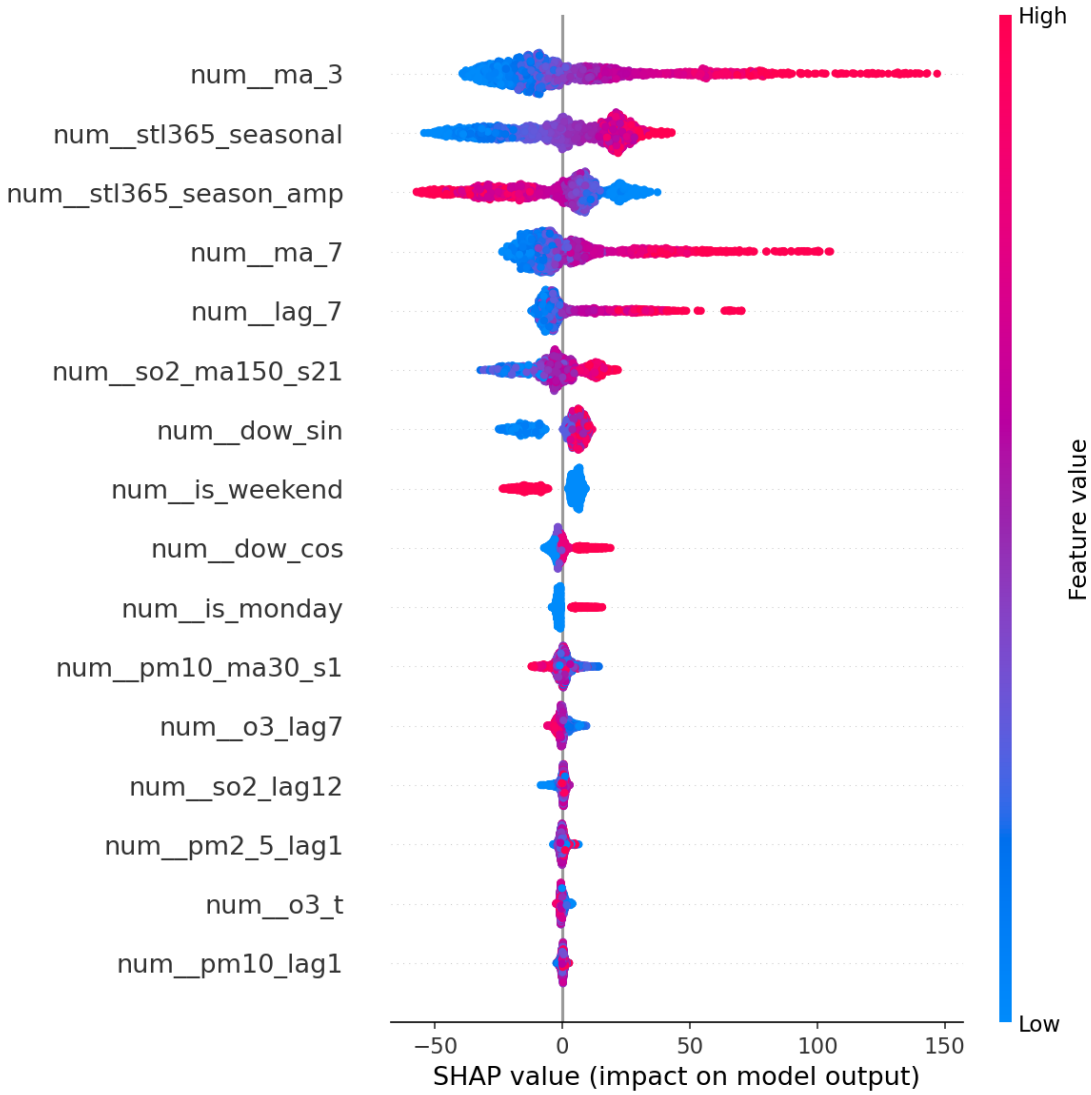

Figura 28: *Summary plot* de SHAP para o RF com *detrending*. As variáveis são ordenadas por impacto médio absoluto na previsão; cores indicam o valor da *feature* (baixo → alto).

Capítulo 6

Conclusões

6.1 Análise Retrospectiva

Este trabalho investigou a viabilidade de prever o número diário de internações por doenças respiratórias no município do Rio de Janeiro a partir da integração de dados públicos de saúde e de variáveis ambientais. Para isso, foi construída uma base diária unificada combinando registros de internações do SIH/SUS (via PySUS) e medições horárias do programa MonitorAr (DATARIO). O fluxo metodológico contemplou etapas de filtragem clínica e geográfica, tratamento de ausências, padronização temporal, transformações estatísticas para reduzir assimetria em variáveis ambientais, análises exploratórias de sazonalidade e dependência temporal, além da engenharia de atributos com *features* autorregressivas, sazonais e ambientais (defasagens e exposição acumulada).

Os resultados confirmaram que a série de internações apresenta componentes temporais relevantes para previsão, incluindo sazonalidade anual e semanal, além de dependência de curto prazo evidenciada por *lags* e médias móveis. Esse resultado serviu como base para a construção de preditores autorregressivos e de calendário, complementados por *features* derivadas de decomposição STL, capazes de representar o nível sazonal esperado e a intensidade do ciclo ao longo do tempo.

Em relação ao algoritmo de aprendizado, o modelo RF foi avaliado em dois cenários: sem *detrending* e com *detrending* causal do alvo. O cenário sem *detrend* apresentou desempenho limitado no conjunto de teste, com viés positivo alto e dificuldade em acompanhar mudanças de patamar no trecho mais recente. Já o cenário com *detrending* causal apresentou melhora significativa e consistente nas métricas e na aderência visual *Real vs. Previsto*, além de reduzir o viés, indicando melhor calibração do nível médio previsto ao longo do tempo. A análise de interpretabilidade via SHAP reforçou que as *features* autorregressivas e sazonais concentram a maioria da informação que ajuda o modelo a prever, enquanto variáveis ambientais tendem a contribuir mais quando representadas por janelas de exposição acumulada e defasagens.

Como limitações, destacam-se: (i) a presença de lacunas e indisponibilidades estruturais em

variáveis ambientais (diferenças de cobertura entre estações e poluentes), exigindo estratégias de imputação e *flags* de ausência; (ii) a alta variabilidade diária das internações, que naturalmente suaviza picos em modelos que aprendem padrões médios; e (iii) a redução do período efetivo de dados para modelagem após a construção de *features* que requerem histórico (por exemplo, decomposição STL em janelas móveis), restringindo o conjunto final a 2014–2024. Ainda assim, o trabalho demonstrou que, com uma construção adequada de base e atributos e com tratamento explícito de tendência, é possível obter previsões mais estáveis e interpretáveis para o problema proposto.

6.2 Trabalhos Futuros

Como continuidade deste estudo, destacam-se três trabalhos que podem ampliar utilidade prática e interpretabilidade do modelo:

- **Avaliar modelos alternativos e previsão probabilística:** comparar o RF com métodos de *gradient boosting* (por exemplo, XGBoost/LightGBM/CatBoost) e/ou abordagens específicas para séries temporais, além de incorporar estimativas de incerteza por meio de intervalos de previsão (quantis ou *conformal prediction*).

- **Aplicar técnicas de pós-processamento para calibração e correção de viés:** investigar estratégias de ajuste das previsões após o treinamento, como correção de viés por regressão sobre o erro (*bias correction*), calibração por quantis (*quantile mapping*) e suavização controlada para reduzir oscilações espúrias sem perder reatividade. O objetivo é aprimorar a aderência *Real vs. Previsto* e reduzir erros sistemáticos em períodos sazonais específicos.

- **Aprimorar avaliação temporal e monitoramento de *drift*:** realizar análises de erro estratificadas por mês/estação do ano e por subperíodos, além de definir um protocolo de re-treinamento periódico e mecanismos de detecção de mudança de distribuição (*concept drift*). Essa linha é particularmente relevante para manter desempenho e calibração quando a série apresenta mudanças estruturais ao longo dos anos.

Referências

Alvarez-Mendoza, C. I., Teodoro, A., Freitas, A., and Fonseca, J. Spatial estimation of chronic respiratory diseases based on machine learning procedures—an approach using remote sensing data and environmental variables in quito, ecuador. *Applied Geography*, 123:102273, 2020. 3.2.1, 3.2.2, 3.2.3, 3.3.3, 3.3.6, 4.3.2

Araujo, L. N., Belotti, J. T., Alves, T. A., Tadano, Y. d. S., and Siqueira, H. Ensemble method based on artificial neural networks to estimate air pollution health risks. *Environmental Modelling and Software*, 123, 2020. 1.1, 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.3.1, 3.3.2, 3.3.3, 3.3.6

Assunção, L., Silva, F., Oliveira, M., and Souza, R. Associations between air quality, meteorological variables, and respiratory disease hospitalizations. *Revista Brasileira de Saúde Ambiental*, 47(2):112–121, 2023. 4.4.1

Barnett-Itzhaki, Z., Nir, V., Kellner, A., Biton, O., Toledano, S., and Klein, A. Machine learning models for predicting pediatric hospitalizations due to air pollution and humidity: A retrospective study. *Pediatric Pulmonology*, 60(5):e71106, 2025. 2.5.1

Buralli, R. J. and Connerton, P. Poluição do ar, saúde e regulação no brasil: estamos avançando? *Cadernos de Saúde Pública*, 2025. 1.1

Cappelli, F., Castronuovo, G., Grimaldi, S., and Telesca, V. Random forest and feature importance measures for discriminating the most influential environmental factors in predicting cardiovascular and respiratory diseases. *International Journal of Environmental Research and Public Health*, 21(7), 2024. 2.5.1

Coelho, F. C., Baron, B. C., de Castro Fonseca, G. M., Reck, P., and Palumbo, D. Alertadengue/pysus: Vaccine, 2021. 4.3.1

Dar, A. A., Jain, A., Malhotra, M., Farooqi, A. R., Albalawi, O., Khan, M. S., and Hiba. Time series analysis with arima for historical stock data and future projections. *Soft Computing*, 28(21):12531–12542, 2024. 2.4.4

Dowlatabadi, Y., abadi, S., Sarkhosh, M., Mohammadi, M., and Moezzi, S. M. M. Assessing the impact of meteorological factors and air pollution on respiratory disease mortality rates: a random forest model analysis (2017–2021). *Scientific Reports*, 14(1), 2024. 1.1

Haben, S., Voss, M., and Holderbaum, W. *Time Series Forecasting: Core Concepts and Definitions*, pages 55–66. Springer International Publishing, Cham, 2023. 2.4.1, 2.4.5

Hyndman, R. J. and Athanasopoulos, G. *Forecasting: Principles and Practice*. OTexts, 2 edition, 2018. 2.4.3

Ji, Y., Zhi, X., Wu, Y., Zhang, Y., Yang, Y., Peng, T., and Ji, L. Regression analysis of air pollution and pediatric respiratory diseases based on interpretable machine learning. *Frontiers in Earth Science*, Volume 11 - 2023, 2023. 2.5.1, 3.2.1, 3.2.2, 3.2.4, 3.3.1, 3.3.3, 3.3.4, 3.3.6, 4.3.2

Joseph, M., Tackes, J., and Bergmeir, C. *Modern Time Series Forecasting with Python: Industry-ready machine learning and deep learning time series analysis with PyTorch and pandas*. Packt Publishing, 2024. 2.4.4

Kachba, Y., de Genaro Chiroli, D. M., Belotti, J. T., Alves, T. A., de Souza Tadano, Y., and Siqueira, H. Artificial neural networks to estimate the influence of vehicular emission variables on morbidity and mortality in the largest metropolis in south america. *Sustainability (Switzerland)*, 12(7), 2020. 1.1, 3.2.1, 3.2.3, 3.3.1, 3.3.2, 3.3.3, 3.3.6

Leites, J., Cerqueira, V., and Soares, C. Lag selection for univariate time series forecasting using deep learning: An empirical study. In Santos, M. F., Machado, J., Novais, P., Cortez, P., and Moreira, P. M., editors, *Progress in Artificial Intelligence*, pages 321–332, Cham. Springer Nature Switzerland, 2025. 2.4.4

Lu, J., Bu, P., Xia, X., Lu, N., Yao, L., and Jiang, H. Feasibility of machine learning methods for predicting hospital emergency room visits for respiratory diseases. *Environmental Science and Pollution Research*, 28(23):29701–29709, 2021. 3.2.1, 3.2.3, 3.3.1, 3.3.2, 3.3.4, 3.3.6

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc., 2017. 2.6

Ministério da Saúde. Poluição atmosférica na ótica do sistema Único de saúde: vigilância em saúde ambiental e qualidade do ar, 2021. 1.1

Miranda, A. C., Santana, J. C. C., Yamamura, C. L. K., Rosa, J. M., Tambourgi, E. B., Ho, L. L., and Berssaneti, F. T. Application of neural network to simulate the behavior of hospitalizations and their costs under the effects of various polluting gases in the city of são paulo. *Air Quality, Atmosphere and Health*, 14(12):2091 – 2099, 2021. 1.1, 3.2.3, 3.3.3, 3.3.6, 4.3.2

Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. O'Reilly Media, Sebastopol, 2019. 2.4.1

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71, 2021. 3.1.4

Prefeitura do Rio de Janeiro. Dados diários do Índice de Qualidade do Ar - IQAR (2017 - 2024), 2024a. 2.3

Prefeitura do Rio de Janeiro. Qualidade do ar - Dados horários (2011 a 2024), 2024b. 4.3.2

Rautela, K. S. and Goyal, M. K. Modelling health implications of extreme pm2.5 concentrations in indian sub-continent: Comprehensive review with longitudinal trends and deep learning predictions. *Technology in Society*, 81, 2025. 1.1

Ravindra, K., Bahadur, S. S., Katoch, V., Bhardwaj, S., Kaur-Sidhu, M., Gupta, M., and Mor, S. Application of machine learning approaches to predict the impact of ambient air pollution on outpatient visits for acute respiratory infections. *Science of The Total Environment*, 858:159509, 2023. 2.5.1

Reis, J. S. d., Costa, R. L., Silva, F. D. d. S., de Souza, E. D. F., Cortes, T. R., Coelho, R. H., Velasco, S. R. M., Neves, D. J. D., Sousa Filho, J. F., Barreto, C. E. C., Cabral Júnior, J. B., dos Reis, H. S., Mendes, K. R., Lins, M. C. C., Ferreira, T. R., Vanderlei, M. H. G. d. S., Alonso, M. F., Mariano, G. L., Gomes, H. B., and Gomes, H. B. Predicting asthma hospitalizations from climate and air pollution data: A machine learning-based approach. *Climate*, 13(2), 2025. 1.1, 3.2.1, 3.2.2, 3.2.3, 3.3.1, 3.3.2, 3.3.3, 3.3.4, 3.3.6, 4.1

Temirbekov, N., Temirbekova, M., Tamabay, D., Kasenov, S., Askarov, S., and Tukenova, Z. Assessment of the negative impact of urban air pollution on population health using machine learning method. *International Journal of Environmental Research and Public Health*, 20(18), 2023. 3.2.1, 3.2.2, 3.2.3, 3.3.2, 3.3.6, 4.1

United States Environmental Protection Agency. AQI Breakpoints | Air Quality System | US EPA, n.d. 2.3, 4.3.2

Widodo, A., Budi, I., and Widjaja, B. Automatic lag selection in time series forecasting using multiple kernel learning. *International Journal of Machine Learning and Cybernetics*, 7(1):95–110, 2016. 2.4.4

World Health Organization. International Classification of Diseases (ICD) — who.int. `https://icd.who.int/browse10/2019/en`, 2019. 2.2

World Health Organization. *World health statistics 2025: monitoring health for the SDGs, Sustainable Development Goals*. World Health Organization, Geneva, 2025. 1.1

Yang, J., Xu, X., Ma, X., Wang, Z., You, Q., Shan, W., Yang, Y., Bo, X., and Yin, C. Application of machine learning to predict hospital visits for respiratory diseases using meteorological and air pollution factors in linyi, china. *Environmental Science and Pollution Research*, 30(38):88431–88443, 2023. 3.2.1, 3.2.2, 3.3.1, 3.3.2, 3.3.4, 3.3.6

Yang, X., Li, Y., Liu, L., and Zang, Z. Prediction of respiratory diseases based on random forest model. *Frontiers in Public Health*, 13, 2025. 2.5.1, 3.2.1, 3.2.3, 3.2.4, 3.3.1, 3.3.3, 3.3.4, 3.3.6, 4.1