Impacto Regional e Demográfico das Doenças Respiratórias nas Dez Capitais Mais Populosas do Brasil em 2020

Diogo Achiles Alves Paz, Enzo da Silva Azevedo

Universidade Tecnológica Federal do Paraná – Campo Mourão (UTFPR-CM)

Departamento de Computação - DACOM

diogopaz@utfpr.edu.br, enzoazevedo@utfpr.edu.br

1. Resumo das Principais Descobertas

Este relatório analisa os impactos regionais e demográficos das doenças respiratórias nas dez capitais brasileiras mais populosas em 2020, a partir dos dados do SIM/DataSUS. Os resultados mostram um aumento acentuado dos óbitos entre março e abril, maior mortalidade absoluta em Curitiba, São Paulo e Recife, e ausência de evidências de que menor escolaridade eleve o risco de morte. Identificou-se ainda que Norte e Nordeste apresentam as maiores proporções de óbitos respiratórios, que mais de 97% das mortes ocorreram com assistência médica e que, na modelagem preditiva, o XGBoost apresentou o melhor desempenho após o ajuste de hiperparâmetros.

2. Definição do Problema e Perguntas de Pesquisa

A pandemia de COVID-19 modificou profundamente o perfil de mortalidade no Brasil em 2020. O aumento expressivo de óbitos por doenças respiratórias motivou este estudo, que busca identificar padrões, desigualdades regionais e grupos mais vulneráveis.

Perguntas de pesquisa:

- A escolaridade influenciou a probabilidade de óbito por doenças respiratórias durante 2020?
- Há diferenças significativas entre as capitais brasileiras na incidência de óbitos por causas respiratórias?

Hipóteses:

- H1: Indivíduos com menor escolaridade apresentaram maior taxa de óbito por causas respiratórias em 2020, nas capitais brasileiras mais populosas.

- H2: Capitais localizadas nas regiões Norte e Nordeste apresentaram maior proporção de óbitos por causas respiratórias em 2020 em comparação às demais regiões.

3. Metodologia Completa e Limitações

A metodologia foi estruturada da coleta à modelagem dos dados. Inicialmente, utilizaram-se pandas e numpy para manipulação dos registros, duckdb para consultas SQL, matplotlib e seaborn para visualização, scipy.stats para testes estatísticos e os modelos LinearRegression, RandomForestRegressor e XGBRegressor para predição, avaliados por MAE, MSE, RMSE e R^2.

A etapa de limpeza envolveu padronização de variáveis, remoção de duplicatas, tratamento de ausências, conversão da idade para anos completos, filtragem das dez capitais mais populosas, identificação dos óbitos respiratórios via CID-10, tratamento de variáveis específicas (como necropsia e investigação), exclusão de registros inconsistentes e mapeamento das capitais para suas regiões.

Para a modelagem, adotou-se a estratégia de separar os meses 1–11 como treino e o mês 12 como teste, permitindo verificar se cada modelo seria capaz de prever corretamente o número de óbitos respiratórios do mês subsequente, conforme as hipóteses definidas.

As análises incluíram consultas SQL, EDA univariada e bivariada, aplicação de testes t e χ^2, criação de variáveis derivadas (feature engineering) e treinamento dos modelos preditivos, seguido pela comparação de desempenho entre eles.

4. Resultados das Análises

4.1 Consultas SQL

4.1.1 Tendência Mensal dos Óbitos Respiratórios

Observou-se uma forte elevação entre março e abril, período do maior impacto inicial da pandemia, meses nos quais se registraram mais de 1000 mortes acima do mês anterior.

4.1.2 Ranking de Capitais por Número de Óbitos

O *ranking* evidenciou Curitiba, São Paulo e Recife como as capitais com maior número de óbitos no ano, e Salvador, Goiânia e Brasília como as com menor incidência de óbitos no mesmo período.

4.1.3 Mortalidade por Escolaridade

Indivíduos com menor nível de escolaridade apresentaram as maiores proporções de óbitos, seguidos pelos de maior nível de escolaridade, sugerindo a ausência de uma relação linear.

4.1.4 Óbitos por Assistência Médica

Mais de 97% dos óbitos respiratórios ocorreram sob assistência médica, o que pode sugerir tanto uma possível saturação do sistema hospitalar quanto a existência de capacidade estrutural suficiente para atender à maioria dos casos que necessitaram de cuidados.

4.1.5 Distribuição Municipal, Escolaridade e Tipo de Doença

Foi possível observar como a mortalidade por doenças respiratórias se distribui entre diferentes capitais e categorias educacionais, fornecendo subsídios para investigar desigualdades regionais e demográficas com maior precisão.

4.2 Análise Exploratória dos Dados (EDA)

4.2.1 Escolaridade

4.2.1.1 Univariada

Os óbitos não respiratórios representam o dobro do número em relação aos respiratórios. No entanto, em uma análise geral, mais de 30% das mortes totais foram causadas por doenças respiratórias, o que constitui um número significativo.

Em relação à distribuição de óbitos por escolaridade, observou-se que esta não apresenta um crescimento linear acompanhando o nível de instrução. No entanto, nos níveis de escolaridade mais alto e mais baixo, é notável uma quantidade de óbitos que, no mínimo, duplica a observada nos demais níveis.

4.1.2 Bivariada

A proporção de óbitos por doenças respiratórias varia moderadamente conforme a escolaridade. As maiores proporções estão nos extremos (sem escolaridade e superior completo), com faixas intermediárias ligeiramente inferiores. Isso indica que o nível educacional não tem uma influência clara ou linear na probabilidade de mortalidade.

4.2 Região

4.2.1 Univariada

As proporções de óbitos por doenças respiratórias variam entre as regiões brasileiras, com valores mais elevados no Sudeste e Nordeste e menores no Norte e Centro-Oeste. A diferença entre as regiões é significativa, sugerindo que o impacto das doenças respiratórias não é uniforme no país.

4.2.2 Bivariada

Em termos absolutos de óbitos nas capitais, como foi possível observar na análise univariada, o Sudeste lidera, seguido por Nordeste e Sul, com o Norte registrando o menor volume. Contudo, em

termos proporcionais, o cenário muda: o Norte, embora tenha o menor número absoluto de casos, apresenta a maior proporção ajustada. Já o Sudeste, mesmo com alto volume bruto, mostra uma das menores proporções. O Nordeste, por sua vez, mantém praticamente a mesma posição tanto em valores absolutos quanto proporcionais.

Esse contraste evidencia que olhar apenas para números absolutos é insuficiente, pois pode mascarar regiões que, proporcionalmente, estão sendo muito mais afetadas, e que, portanto, podem demandar maior atenção e apoio.

5. Testes de Hipóteses

5.1 Hipótese H1 — Escolaridade

- Teste t unicaudal à direita
- Estatística: t = –3,6953
- p-valor: 0,999890

Conclusão: não há evidências de que menor escolaridade aumente o risco de morte por doença respiratória.

5.2 Hipótese H2 — Região

- Estatística t: 16,3920
- p-valor ≈ 0

Conclusão: Norte e Nordeste apresentam proporções significativamente mais altas de óbitos respiratórios.

6. Modelagem Preditiva

6.1 Dataset

- 59.176 registros
- 16 variáveis

6.2 Pré-Visualizações

As visualizações exploratórias foram utilizadas para comparar com o comportamento da regressão linear ao longo do tempo e identificar se havia tendência de crescimento linear nas análises.

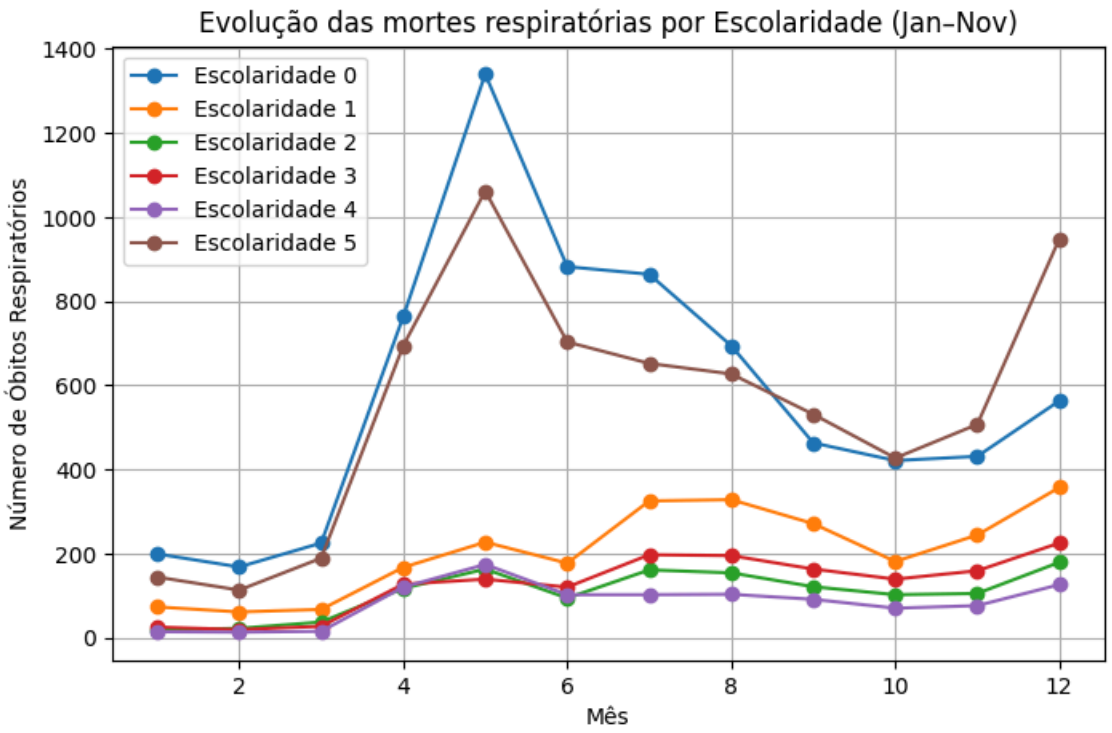

Figura 1. Evolução das mortes respiratórias por escolaridade.

Fonte: Elaboração própria (2025)

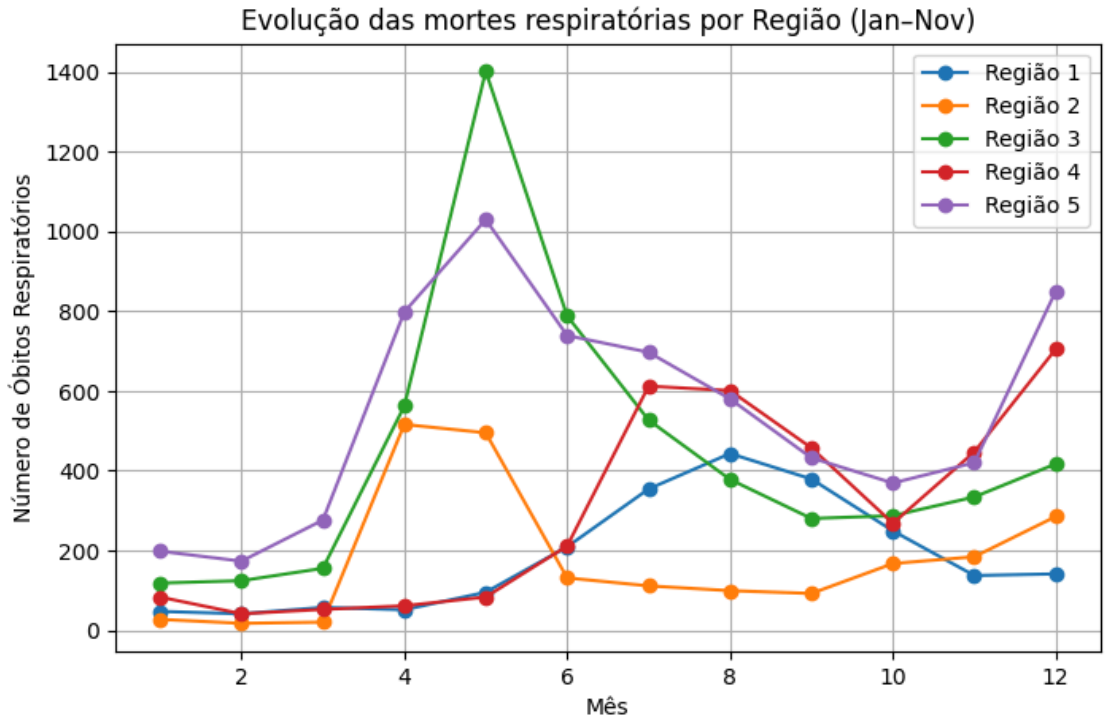

Figura 2. Evolução das mortes respiratórias por região.

Fonte: Elaboração própria (2025)

6.3 Feature Engineering

Inclui:

- Mapeamento numérico das regiões;
- Extração do mês do óbito;
- Criação de faixas etárias decenais;
- Conversão de variáveis categóricas em numéricas;
- Definição da variável alvo.

6.4 Comparação Entre Modelos

Primeira execução:

	Modelo	MAE	RMSE	R^2
2	XGBoost	27.514692	60.057888	0.704361
1	Random Forest	27.936063	61.070448	0.694308
0	Linear Regression	31.346915	57.871630	0.725493

Figura 3. Primeira Execução dos Modelos Preditivos.

Fonte: Elaboração própria (2025)

Execução tunada:

	Modelo	MAE	RMSE	R^2
2	XGBoost	26.585915	47.714434	0.813396
1	Random Forest	26.868942	57.433557	0.729634
0	Linear Regression	31.346915	57.871630	0.725493

Figura 4. Execução Tunada dos Modelos Preditivos.

Fonte: Elaboração própria (2025)

Hiperparâmetros ajustados:

Modelo	Hiperparâmetro	Inicial	Tunado
XGBoost	n_estimators	300	100
	learning_rate	0.1	0.10
	max_depth	5	9
	subsample	0.8	0.31
	colsample_bytree	0.8	0.65
Random Forest	n_estimators	300	7000
	max_depth	6	16

Tabela 1. Comparação dos Hiperparâmetros Antes e Depois.

Fonte: Elaboração própria (2025)

8. Discussão

O estudo revela que a pandemia afetou desproporcionalmente as regiões Norte e Nordeste, refletindo desigualdades estruturais. A escolaridade não mostrou impacto estatisticamente significativo, embora diferenças numéricas indiquem possíveis efeitos indiretos.

A forte elevação no segundo trimestre de 2020 reforça o papel central da COVID-19 na mortalidade respiratória. O elevado percentual de óbitos com assistência médica pode indicar sobrecarga do sistema de saúde, possivelmente associado à superlotação de hospitais e UTIs.

9. Recomendações

Recomenda-se ampliar a infraestrutura de saúde nas regiões mais vulneráveis, aprimorar a qualidade e completude dos dados do SIM e adotar estratégias de prevenção adaptadas ao contexto epidemiológico de cada região. Sugere-se também integrar técnicas de modelagem preditiva à vigilância em saúde e fortalecer políticas voltadas à redução de desigualdades estruturais que influenciam a mortalidade por doenças respiratórias.

10. Trabalhos Futuros

Trabalhos futuros podem incluir a extensão da análise para o período de 2021 a 2023, permitindo observar o restante do ciclo pandêmico e identificar possíveis mudanças no comportamento da mortalidade respiratória. Também é recomendada a realização de avaliações específicas por faixa etária e por sexo, aprofundando a compreensão das diferenças demográficas. Além disso, a aplicação de técnicas de clusterização, como o K-Means, pode auxiliar na identificação de perfis

distintos de óbitos a partir de padrões multidimensionais envolvendo variáveis como idade, local do óbito, mês e região.

11. Referências

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Análise em Saúde e Vigilância de Doenças Não Transmissíveis. Dicionário de dados do Sistema de Informações sobre Mortalidade (SIM). Arquivo atualizado em jun. 2025. Disponível em: https://s3.sa-east-1.amazonaws.com/ckan.saude.gov.br/SIM/Dicionario_SIM_2025.pdf. Acesso em: 05 nov. 2025.

BRASIL. Dados abertos: SIM – Sistema de Informações sobre Mortalidade (1979-2019). Dados.gov.br. Disponível em: https://dados.gov.br/dados/conjuntos-dados/sim-1979-2019. Acesso em: 05 nov. 2025.

PENA, E. Materiais da disciplina Ciência de Dados, disponibilizados no ambiente Moodle. Acesso em: 10 nov. 2025.