Análise de Fatores de Óbito por COVID-19 no Brasil utilizando Regressão Logística

Danielle Ester Barbosa da Silva[1] e Luísa Oliveira Gonçalves[2]

Instituto Federal de Brasília, Brasil

Resumo A pandemia de COVID-19 impôs desafios significativos aos sistemas de saúde em escala global, demandando o uso de métodos computacionais capazes de apoiar a tomada de decisão clínica e a formulação de políticas públicas. Este estudo apresenta uma análise exploratória e preditiva de dados epidemiológicos de Síndrome Respiratória Aguda Grave (SRAG) associados à COVID-19 no Brasil, utilizando registros públicos do Ministério da Saúde referentes ao período de 2019 a 2025, totalizando mais de quatro milhões de casos. Inicialmente, foi conduzido um processo sistemático de preparação e limpeza dos dados, seguido por uma análise exploratória voltada à identificação da distribuição regional de comorbidades. Posteriormente, foi desenvolvido um modelo de regressão logística para estimar a probabilidade de óbito, considerando variáveis demográficas, clínicas e regionais. O modelo apresentou desempenho satisfatório, com AUC-ROC de 0,7694, indicando boa capacidade discriminativa. Os resultados evidenciam diferenças regionais relevantes e confirmam a importância de fatores como idade, internação em UTI e presença de comorbidades, especialmente cardiopatias. Conclui-se que a aplicação de técnicas de ciência de dados em bases epidemiológicas nacionais pode contribuir de forma significativa para a compreensão dos fatores associados à gravidade da COVID-19 no Brasil.
Palavras-chave: COVID-19; Regressão Logística; Análise Regional

1 Introdução

Com os primeiros casos confirmados no ano de 2019, a infecção pelo vírus SARS-CoV-2 rapidamente se consolidou como uma emergência de saúde pública global, acumulando mais de 700 milhões de casos confirmados em todo o mundo, segundo dados da Organização Mundial da Saúde [1]. No Brasil, a pandemia evidenciou desigualdades regionais no acesso aos serviços de saúde e na capacidade de resposta dos sistemas locais, o que reforçou a necessidade de monitoramento epidemiológico contínuo e de análises baseadas em dados.

Nesse contexto, o Ministério da Saúde ampliou e consolidou a coleta de registros relacionados à Síndrome Respiratória Aguda Grave, disponibilizando bases de dados públicas por meio da plataforma OpenDataSUS [2]. Esses registros possibilitam análises retrospectivas e prospectivas sobre a evolução da COVID-19, bem como a identificação de fatores associados a desfechos clínicos graves, como o óbito.

Técnicas de aprendizado de máquina têm sido amplamente empregadas na área da saúde para identificação de padrões complexos e apoio à tomada de decisão clínica, especialmente em cenários que envolvem grandes volumes de dados [3]. A utilização desses métodos pode contribuir para a redução de custos operacionais, otimização de recursos e melhoria da qualidade do cuidado em saúde [4]. Diante desse cenário, o presente trabalho tem como objetivo realizar uma análise exploratória e desenvolver um modelo preditivo para compreender os fatores associados ao óbito por COVID-19 no Brasil, com ênfase nas diferenças regionais observadas entre os anos de 2019 e 2025.

2 Materiais e Métodos

Este estudo utilizou dados epidemiológicos públicos provenientes dos sistemas de informação do Ministério da Saúde, disponibilizados por meio da plataforma OpenDataSUS. Foram consideradas bases consolidadas referentes ao período de 2019 a 2025, compostas por milhões de registros de casos de Síndrome Respiratória Aguda Grave associados à COVID-19. A seleção das variáveis foi orientada pelo dicionário de dados oficial, priorizando informações demográficas, regionais e clínicas relevantes para a análise do desfecho óbito.

O processo de preparação dos dados envolveu a filtragem dos registros com evolução clínica conhecida, mantendo apenas os casos classificados como cura ou óbito. A variável alvo foi definida de forma binária, sendo atribuído o valor zero aos casos de cura e o valor um aos casos de óbito. A idade dos pacientes foi convertida para formato numérico contínuo, enquanto a unidade federativa de residência foi mapeada para as cinco grandes regiões geográficas do Brasil. Foram selecionadas como variáveis explicativas a idade, a necessidade de internação hospitalar, a internação em unidade de terapia intensiva e a presença de comorbidades específicas, incluindo diabetes, cardiopatias, obesidade, pneumopatias, doenças neurológicas, renais e imunodepressão.

As variáveis clínicas binárias foram recodificadas para formato numérico, considerando apenas respostas válidas, e registros com valores ausentes após o processo de recodificação foram removidos. Essas etapas asseguraram a consistência e a qualidade do conjunto final utilizado na modelagem preditiva.

Após o pré-processamento, foi realizada uma análise exploratória descritiva com foco na distribuição regional das comorbidades mais frequentes entre os pacientes diagnosticados com COVID-19. Para essa finalidade, foi gerado um gráfico do tipo *heatmap*, permitindo a visualização comparativa das principais comorbidades entre as cinco regiões do país. Foram consideradas as comorbidades mais recorrentes no conjunto de dados, destacando-se cardiopatias, diabetes, obesidade, pneumopatias e doenças neurológicas.

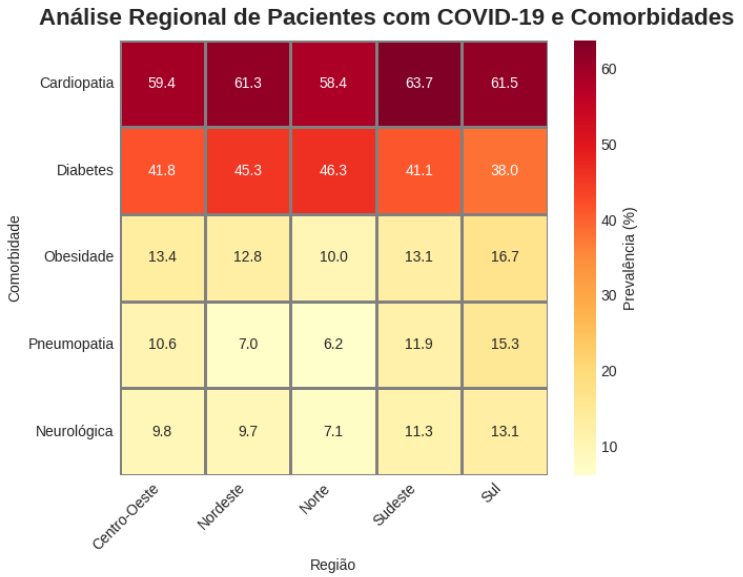

Fig. 1. Distribuição regional das principais comorbidades em pacientes com COVID-19

Para a análise preditiva do desfecho óbito, foi empregado um modelo de regressão logística, implementado por meio de um pipeline que integrou as etapas de pré-processamento e aprendizado. A variável idade foi padronizada utilizando normalização, enquanto a variável categórica referente à região foi transformada por meio de codificação *one-hot*, com remoção da categoria de referência. As variáveis binárias foram mantidas em seu formato original após a recodificação.

O conjunto de dados foi dividido em subconjuntos de treinamento e teste, utilizando-se 70% dos registros para treinamento e 30% para teste, com estratificação da variável alvo a fim de preservar a proporção entre os desfechos. Considerando o desbalanceamento entre as classes de cura e óbito, foi adotado o balanceamento de classes durante o treinamento do modelo.

A avaliação do desempenho foi realizada por meio da métrica de área sob a curva ROC (AUC-ROC), calculada no conjunto de teste. Adicionalmente, foi aplicada validação cruzada com cinco partições sobre o conjunto de treinamento, utilizando a AUC-ROC como métrica de avaliação, com o objetivo de analisar a estabilidade e a capacidade de generalização do modelo.

3 Resultados

O modelo de regressão logística apresentou desempenho satisfatório na tarefa de discriminação entre os desfechos de cura e óbito por COVID-19. A Figura 2 apresenta a curva ROC (Receiver Operating Characteristic) do modelo de

regressão logística utilizado para a predição do desfecho analisado. A curva ROC representa a relação entre a taxa de verdadeiros positivos (sensibilidade) e a taxa de falsos positivos (1 – especificidade) para diferentes limiares de decisão do modelo.

Observa-se que a curva do modelo permanece consistentemente acima da diagonal que representa um classificador aleatório, indicando que o modelo possui capacidade discriminativa superior ao acaso. O valor da AUC-ROC obtido foi de 0,7694, o que caracteriza um desempenho considerado satisfatório, evidenciando que o modelo consegue distinguir adequadamente entre os indivíduos que evoluíram para óbito e aqueles que apresentaram cura.

Adicionalmente, a validação cruzada resultou em uma AUC média de 0,7702 ± 0,0008, demonstrando estabilidade e consistência do desempenho do modelo entre diferentes subconjuntos dos dados. Esses resultados indicam que a regressão logística apresenta bom poder preditivo para o problema em estudo.

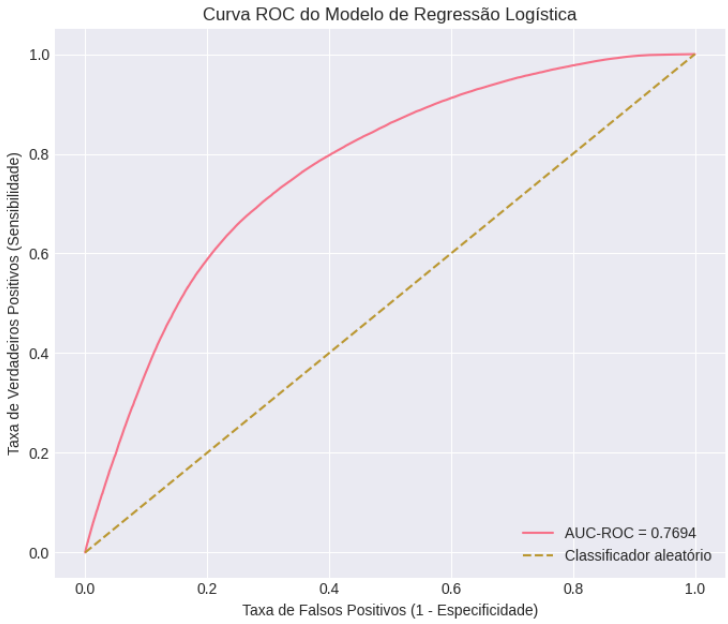

Fig. 2. Curva ROC da Regressão Logística

A regressão logística também possibilita compreender quais fatores estão mais associados ao desfecho analisado, e não apenas realizar previsões. Esse aspecto é especialmente importante na área da saúde pública, pois permite identificar fatores de risco relevantes para apoiar decisões e ações preventivas. Os resultados mostram que variáveis clínicas, como a necessidade de internação hospitalar e de internação em unidade de terapia intensiva, assim como a pre-

sença de comorbidades, estão fortemente associadas ao óbito. Além disso, foram observadas diferenças entre regiões do país, e a idade permaneceu como um fator importante, estando associada ao aumento da probabilidade de óbito.

4 Considerações Finais

Os resultados obtidos neste estudo evidenciam o potencial da aplicação de técnicas de ciência de dados e aprendizado de máquina na análise de grandes bases epidemiológicas, contribuindo para a compreensão dos fatores associados à gravidade da COVID-19 no Brasil. O desempenho alcançado pelo modelo de regressão logística, com AUC-ROC de 0,7694, pode ser considerado relevante tanto do ponto de vista técnico quanto sob a perspectiva da aplicabilidade prática em saúde pública.

Do ponto de vista da aplicação prática, o modelo foi treinado e validado a partir de um volume expressivo de registros, conferindo maior confiabilidade estatística aos achados. A utilização de variáveis coletadas no contexto hospitalar, como idade, presença de comorbidades e informações sobre internação, torna a abordagem viável para implementação em larga escala, sem custos adicionais associados à obtenção de novos dados. Nesse contexto, o modelo pode atuar como ferramenta de apoio à triagem e à priorização de pacientes, auxiliando na alocação de recursos e no planejamento de ações em cenários de alta demanda assistencial.

A análise também evidenciou diferenças regionais relevantes nos padrões de risco, reforçando a importância de estratégias regionalizadas no enfrentamento da COVID-19. Em um país marcado por desigualdades demográficas e estruturais, a identificação e a quantificação dessas diferenças constituem subsídios importantes para a formulação de políticas públicas mais equitativas e eficazes.

Apesar dos resultados, é necessário reconhecer limitações inerentes ao estudo, especialmente aquelas relacionadas à qualidade dos dados secundários utilizados, à possibilidade de subnotificação e à ausência de variáveis socioeconômicas mais detalhadas. Além disso, a natureza observacional dos dados impede inferências causais diretas, e mudanças nos protocolos de tratamento ao longo do período analisado podem influenciar o resultado do modelo.

Como perspectivas futuras, estudos adicionais podem incorporar variáveis socioeconômicas e variantes virais, com a comparação com abordagens baseadas em algoritmos mais complexos. Por fim, os resultados apresentados demonstram que a combinação de dados públicos em larga escala com metodologias estatisticamente sólidas pode gerar conhecimento, contribuindo para o fortalecimento da vigilância epidemiológica baseada em dados no Brasil.

References

1. World Health Organization: Coronavirus disease (COVID-19) pandemic. https://www.who.int/emergencies/diseases/novel-coronavirus-2019, último acesso em 10-12-2025.

2. Ministério da Saúde: Banco de dados da Síndrome Respiratória Aguda Grave (SRAG) – 2019 a 2025. https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024, último acesso em 10-12-2025.
3. Kietzmann, T.: Avaliação de modelos preditivos de aprendizado de máquina como suporte na tomada de decisão gerencial: a predição de risco de mortalidade por COVID-19 no estado de São Paulo. https://www.teses.usp.br, último acesso em 16-12-2025.
4. Baron, J. et al.: Use of machine learning to predict clinical decision support compliance, reduce alert burden, and evaluate duplicate laboratory test ordering alerts. https://pmc.ncbi.nlm.nih.gov/articles/PMC7935497/, último acesso em 16-12-2025.
5. GRUS, J.: Data science do zero: noções fundamentais com Python. 2. ed. Rio de Janeiro: Alta Books, 2016. Livro digital.. ISBN 9788550816463. https://integrada.minhabiblioteca.com.br/books/9788550816463. último acesso em: 17-12-2025.