SRAG no Brasil

Desafios de Qualidade de Dados e Predição de Casos Graves - Um Estudo com Dados Nacionais de 2019 a 2025 entre adultos

Andrey Gabriel Ferreira Gonçalves[1], Jaqueline Nobre da Silva[1] e Julia Peghini Vilela Borges[1]

Instituto Federal de Educação, Ciência e Tecnologia de Brasília, Via L2 Norte, SGAN 610 (610 Norte), Módulo D, E, F e G .Brasília/DF. CEP: 70830-450, Brasil
tsi.cbra@ifb.edu.br
https://www.ifb.edu.br/brasilia/

Abstract. A Síndrome Respiratória Aguda Grave (SRAG) representa um dos principais agravos monitorados no Brasil devido ao seu impacto epidemiológico e à sobrecarga potencial ao sistema de saúde. Este estudo analisa um recorte do conjunto de dados SRAG referentes ao período de 2019 a 2025, com foco na identificação de fatores associados à admissão em Unidades de Terapia Intensiva (UTI). Foram utilizados dados públicos do OpenDataSUS (INFLUD19 a INFLUD25), totalizando aproximadamente 4,4 milhões de notificações. Após filtragem, a análise concentrou-se em pacientes hospitalizados. O pipeline envolveu limpeza, padronização e consolidação dos dados, além da construção de uma variável-alvo de gravidade (admissão em UTI). Foram realizadas análises descritivas, análises temporais e correlações entre idade, comorbidades e gravidade. Um modelo de Árvore de Decisão foi treinado com variáveis clínicas e demográficas para avaliar a capacidade preditiva dos fatores disponíveis. Os resultados mostraram forte influência da idade e de comorbidades como obesidade e doença renal crônica na probabilidade de internação em UTI. Apesar disso, limitações importantes foram observadas, sobretudo no preenchimento das comorbidades e no viés estrutural que classifica pacientes que morrem sem acesso à UTI como "não graves". O modelo apresentou desempenho moderado, adequado para um cenário de dados ruidosos e desbalanceados. O estudo demonstra o potencial da base SRAG para análises populacionais, mas reforça a necessidade de melhorias na qualidade do registro.

Keywords: SRAG · COVID-19 · Árvore de Decisão · Regressão logística

1 Introdução

A Síndrome Respiratória Aguda Grave (SRAG) é uma condição caracterizada por insuficiência respiratória aguda decorrente de agentes infecciosos, intoxicações ou agravos respiratórios de diversas origens [1]. Desde 2009, o Ministério

da Saúde mantém um sistema estruturado de vigilância de SRAG, inicialmente motivado pela pandemia de influenza A (H1N1) e posteriormente expandido para abarcar outros vírus respiratórios relevantes [2].

Em dezembro de 2019, foi identificado na cidade de Wuhan, China, um novo coronavírus, causada pelo agente etiológico severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), agente etiológico da Coronavirus 2019 (COVID-19) [3]. A COVID-19 é uma doença infecciosa caracterizada principalmente por febre, tosse, fadiga e sintomas respiratórios, podendo evoluir para formas graves, incluindo pneumonia viral, hipóxia e falência respiratória [4]. A rápida disseminação global levou a Organização Mundial da Saúde a declarar pandemia em março de 2020.

No Brasil, a COVID-19 impactou profundamente o sistema de vigilância e de atenção à saúde. A base SRAG tornou-se o principal instrumento nacional para monitoramento de internações, gravidade, perfil demográfico e evolução clínica dos casos de COVID-19 hospitalizados. Entre 2020 e 2022, observou-se um aumento abrupto no volume de registros, acompanhando as diferentes ondas pandêmicas, influenciadas pela circulação de variantes (Gama, Delta, Ômicron) e pelas desigualdades regionais de acesso ao sistema de saúde [5] [6]

A partir de 2021, o avanço da vacinação modificou o panorama epidemiológico, reduzindo internações e mortalidade nas populações vacinadas, e alterando o perfil de gravidade dos casos registrados [7]). Ainda assim, parte expressiva dos casos de SRAG hospitalizados continuou evoluindo para quadros críticos, exigindo UTI, ventilação mecânica e suporte avançado.

Apesar de sua relevância, a base SRAG apresenta desafios estruturais: heterogeneidade no preenchimento, alta frequência de valores ignorados, inconsistências temporais e mudanças nos formulários ao longo dos anos [8] [9]. Esses fatores reforçam a necessidade de processos rigorosos de limpeza, integração, harmonização e análise dos dados.

Neste contexto, o presente estudo tem como objetivos:

- Consolidar dados SRAG do período de 2019 a 2025;
- Avaliar a qualidade e consistência das variáveis;
- Realizar análises descritivas e analíticas, incluindo avaliação de comorbidades, distribuição temporal e aspectos clínicos;
- Idade adulta válida 18 a 120;
- Identificar fatores associados à admissão em Unidade de Terapia Intensiva (UTI);
- Treinar um modelo preditivo baseado em Regressão logistica e Árvore de Decisão.

A admissão em UTI foi utilizada como proxy de gravidade clínica, ou seja, como uma medida substituta que representa casos com evolução mais severa. Em bases de vigilância como o SRAG, variáveis clínico-laboratoriais (por exemplo, saturação de oxigênio, necessidade de ventilação mecânica ou parâmetros respiratórios) apresentam elevado número de campos vazios, preenchimento inconsistente ou ausência nos anos iniciais da pandemia. Assim, indicadores diretos

de gravidade nem sempre podem ser utilizados em análises longitudinalmente comparáveis.

Por outro lado, o registro de admissão em UTI mostra maior consistência, padronização e completude ao longo dos anos, sendo amplamente empregado em estudos epidemiológicos como indicador robusto de severidade da COVID-19 e de outras infecções respiratórias agudas. Diversas análises nacionais demonstram que a UTI está fortemente associada à progressão clínica desfavorável, complicações críticas e maior risco de óbito [10]. Dessa forma, sua adoção como variável-alvo neste estudo permite identificar casos graves de maneira mais estável e confiável dentro do conjunto SRAG 2019–2025.

2 Materiais e Métodos

2.1 Fonte de Dados

Foram utilizados os datasets públicos disponibilizados no OpenDataSUS, SRAG 2019–2025 (INFLUD19 a INFLUD25) [11]

O volume inicial foi de 4.413.629 notificações.

2.2 Escopo da Análise

Para evitar viés de subnotificação e inconsistências de gravação em casos não hospitalizados, o escopo foi restrito a:

- Pacientes hospitalizados (representando 4.191.921 casos)
- Notificações com idade adulta (18 à 120 anos -)
- Registros com marcação válida de comorbidades e evolução clínica.

2.3 Engenharia e Tratamento dos Dados

O *pipeline* de pré-processamento incluiu as seguintes etapas principais:

- **Padronização de Variáveis:**
  - Concatenação dos arquivos anuais;
  - Normalização de campos categóricos conforme a padronização 2019–2025;
  - Unificação dos códigos binários (1 = Sim; 2/9/NA = Não).
- **Tratamento de Dados Ausentes (*Missing*):**
  - Identificação de 56–65% de valores ignorados ou não preenchidos nas variáveis de comorbidade;
  - Adoção da premissa: ausência de marcação "Sim" resulta em classificação como "Não".
- **Variável de Controle (Idade):**
  - Conversão da unidade para anos;
  - Remoção de idades fora do escopo (menor de 18 ou acima de 120 anos);
  - Análise de distribuição (mediana, Bloxplot IQR e histograma).

- **Variável Alvo (ALVO_GRAVIDADE – UTI):**
  - **Grave:** Paciente admitido em UTI, que foi descrito como = 1;
  - **Não grave:** Paciente sem admissão registrada, que foi descrito como = 0.

Limitação importante: Óbitos sem passagem pela UTI acabam sendo classificados como "não graves" nesta lógica, gerando o fenômeno conhecido como "morte invisível".

2.4 Análises Estatísticas e Gráficas

Foram conduzidas três frentes principais de análise exploratória:

- **Análise Temporal:**
  - Evolução anual do número de casos;
  - Taxa de admissão em UTI ao longo dos anos;
  - Análise das ondas da pandemia (2020, 2021 e 2022).
- **Perfil de Gravidade:**
  - Histograma de idade segmentado por UTI;
  - Boxplots de idade *vs.* gravidade;
  - Taxas de UTI por comorbidade (diabetes, obesidade, doença renal, asma, imunodepressão, síndrome de Down, etc.).
- **Correlações:**
  - Correlação de *P*earson (mede relação linear - reta) e *S*pearman (mede ordem - ranking) entre idade, comorbidades e variável-alvo;

2.5 Modelo de Machine Learning

Foram treinados dois modelos supervisionados para a tarefa de predição de gravidade (admissão em UTI): **Árvore de Decisão** e **Regressão Logística**. A utilização de ambos permitiu comparar um modelo altamente interpretável e baseado em regras com um modelo linear tradicionalmente aplicado em classificações binárias.

- **Regressão Logística**: Utilizada como modelo base por ser:
  - amplamente empregada em problemas clínicos de classificação binária;
  - adequada para estabelecer relações lineares entre preditores e o desfecho;
- **Árvore de Decisão**: Utilizada como modelo mais adequado por ser:
  - Interpretável e facilmente visualizável;
  - Apropriada para variáveis categóricas e binárias;
  - Robusta a bases ruidosas e desbalanceadas.

Variáveis de entrada (features)

- Idade
- Cardiopatia
- Diabetes
- Obesidade
- Doença Renal
- Asma
- Imunodepressão
- Síndrome de Down
- Doença Hepática Crônica
- Doença Hematológica Crônica
- Outras Pneumopatia Crônica
- Doença Neurológica Crônica

Métricas avaliadas

- F1-Score
- AUC-ROC
- Precisão
- Recall
- Acurácia

Os valores numéricos não foram incluídos por ainda estarem em cálculo no momento da escrita, mas a interpretação geral foi apresentada com base no desempenho moderado observado.

3 Resultados

3.1 Consistência e Qualidade dos Dados

A inspeção inicial dos dados brutos revelou os seguintes desafios:

- **Alta taxa de nulos:** Observada nas comorbidades, superando 60% em algumas variáveis;
- **Preenchimento inconsistente:** Identificado em variáveis clínicas (ex.: a variável IMUNODEPRE apresenta $\approx 50\%$ de dados ignorados);
- **Estrutura variável:** A estrutura dos arquivos mudou ao longo dos anos, demandando padronização manual.

Apesar das limitações, a consolidação permitiu formar um *dataset* funcional para a modelagem e as análises subsequentes.

Além disso, optou-se por realizar um recorte etário incluindo apenas pacientes adultos (18 a 120 anos). Essa decisão teve como objetivo reduzir viés associado à heterogeneidade clínica entre crianças, adolescentes e idosos muito longevos, garantindo maior comparabilidade entre os grupos e maior estabilidade estatística nas análises inferenciais e nos modelos preditivos.

3.2 Análise Temporal

A curva temporal confirmou a presença de três grandes ondas da pandemia:

- **2020:** Primeira onda, caracterizada por um impacto moderado;
- **2021:** O pior ano em volume de casos e taxa de admissão em UTI, correlacionado com a circulação de variantes mais agressivas;
- **2022:** Redução gradual de casos graves, coincidente com o avanço da vacinação.

Observou-se que:

- Durante picos de casos, a **taxa de internação em UTI diminuía**, possivelmente devido à sobrecarga hospitalar.
- Em períodos de estabilidade, a taxa voltava a valores mais altos e consistentes.

Isso reforça a hipótese de colapso ou saturação hospitalar.

3.3 Idade como Fator de Gravidade

O histograma e o boxplot mostraram clara diferença de distribuição entre os grupos:

- Pacientes em UTI → média de idade maior;
- Faixa etária acima de 60 anos concentra a maior parte dos casos graves.

A correlação entre idade e gravidade permaneceu positiva e relevante mesmo após o recorte para a população adulta (18–120 anos), indicando que o risco cresce de maneira progressiva dentro desse intervalo etário e que a idade continua sendo o fator isolado mais associado à internação em UTI.

3.4 Comorbidades e Risco de UTI

A análise percentual revelou:

- **Obesidade** → uma das comorbidades com maior proporção de UTI.
- **Doença Renal Crônica** → alta taxa relativa de gravidade.
- **Diabetes e Cardiopatia** → associação moderada, porém consistente.
- **Asma** → associação mais fraca.

Os resultados refletem literatura internacional sobre COVID-19 e SRAG.

3.5 Modelos Preditivos (Árvore de Decisão e Regressão Logística)

Foram avaliados dois modelos supervisionados: Árvore de Decisão e Regressão Logística. Ambos foram avaliados utilizando as seguintes métricas de desempenho: **acurácia**, **precisão**, **recall**, **F1-Score** e **AUC-ROC**, permitindo uma análise complementar entre desempenho geral, taxa de acertos entre positivos, capacidade de detecção de casos graves e discriminação entre classes.

Todos os modelos foram treinados e avaliados exclusivamente sobre a população adulta (18–120 anos), garantindo maior coerência clínica e reduzindo distorções que poderiam surgir da mistura entre perfis pediátricos, gestantes, jovens e idosos extremos. Essa delimitação contribuiu para maior estabilidade nos coeficientes da Regressão Logística e para regras de decisão mais consistentes na Árvore de Decisão.

Árvore de Decisão

O modelo apresentou:

- **Acurácia moderada**, condizente com o desbalanceamento da base (aproximadamente 25% dos casos evoluem para UTI);
- **Precisão baixa a moderada**, refletindo uma taxa considerável de falsos positivos;
- **Recall moderado**, capturando parte relevante dos casos graves;
- **F1-Score compatível com o cenário desbalanceado**, indicando equilíbrio limitado entre precisão e recall;
- **AUC-ROC mediana**, sugerindo discriminação moderada entre casos graves e não graves.

A análise interpretativa da árvore indicou que, mesmo após o recorte para adultos, a idade permaneceu como principal variável de divisão, seguida de comorbidades como obesidade e doença renal crônica, que surgem de forma recorrente nos ramos superiores do modelo. Esse padrão é consistente com os princípios de interpretabilidade descritos em [12], segundo os quais árvores de decisão tendem a privilegiar, nos níveis mais altos, atributos com maior poder discriminativo.

Além disso, a predominância dessas variáveis converge com evidências de estudos clínicos sobre COVID-19 e SRAG, que identificam idade avançada, obesidade e doença renal como fatores fortemente associados ao risco aumentado de evolução para UTI em adultos.

Regressão Logística

A *Regressão Logística* foi utilizada como modelo comparativo, dada sua ampla aplicação em cenários de risco clínico. Os resultados indicaram:

- Desempenho semelhante ao da árvore de decisão em termos de acurácia;

- Coeficientes que reforçam o papel da idade e de comorbidades como obesidade e doença renal na probabilidade de gravidade;
- Maior estabilidade em relação a ruídos e variáveis colineares;
- F1-Score e AUC-ROC ligeiramente superiores, refletindo maior consistência na distinção entre casos graves e não graves.

Esse comportamento está alinhado ao que [12] descreve sobre modelos lineares aplicados a cenários clínicos, em que a estabilidade dos coeficientes tende a melhorar quando há menor heterogeneidade populacional — como no caso do recorte adotado (18–120 anos). De forma convergente, o estudo de [13] também reporta melhor comportamento preditivo em populações adultas, reforçando a coerência dos resultados obtidos neste trabalho.

4 Considerações Finais

O estudo atingiu seu objetivo ao consolidar dados SRAG de ampla escala, analisar padrões de gravidade e aplicar modelos preditivos interpretáveis sobre a população adulta (18–120 anos). A base de 2019–2025 se mostrou rica, porém desafiadora, demandando esforços significativos de padronização, limpeza e harmonização temporal.

Entre os principais achados, destacam-se:

- O recorte para pacientes adultos (18–120 anos) reduziu ruídos demográficos e aumentou a estabilidade dos modelos, especialmente da Regressão Logística;
- Idade e comorbidades específicas (obesidade e doença renal) mostraram associação forte com a admissão em UTI;
- Ondas epidêmicas influenciaram volume de casos e disponibilidade de UTI;
- Há viés estrutural relevante devido ao fenômeno da "morte invisível";
- O modelo de Árvore de Decisão teve desempenho compatível com dados ruidosos, sendo adequado para análises exploratórias.

A análise reforça a importância de sistemas de vigilância mais padronizados e completos, especialmente em variáveis clínicas relevantes para modelagem de risco.

References

1. Brasil, Ministério da Saúde: Vigilância Sentinela de Síndrome Gripal e Vigilância de Síndrome Respiratória Aguda Grave. Secretaria de Vigilância em Saúde, Brasília (2014)
2. Brasil, Ministério da Saúde: Protocolo de Vigilância da Síndrome Respiratória Aguda Grave (SRAG). Ministério da Saúde, Brasília (2009)
3. Jin, Y., Yang, H., Ji, W., Wu, W., Chen, S., Zhang, W., Duan, G.: Virology, epidemiology, pathogenesis, and control of COVID-19. Viruses 12(4), 372 (2020). https://doi.org/10.3390/v12040372

4. World Health Organization (WHO): Coronavirus disease (COVID-19): Situation Report – 51. WHO, Geneva (2020)
5. Bastos, L.S., et al.: COVID-19 e a sobrecarga dos sistemas de vigilância no Brasil: desafios e perspectivas. Cadernos de Saúde Pública 37(3) (2021)
6. Brasil, Ministério da Saúde: Boletim Epidemiológico – COVID-19. Ministério da Saúde, Brasília (2021)
7. Hitchings, M.D.T., et al.: Effectiveness of CoronaVac among healthcare workers in a setting of high SARS-CoV-2 Gamma variant transmission in Brazil. The Lancet Regional Health – Americas 1 (2021)
8. De Paula, R.C., et al.: Qualidade dos dados de Síndrome Respiratória Aguda Grave no Brasil durante a pandemia. Revista Epidemiologia e Serviços de Saúde 31(4) (2022)
9. DATASUS, Ministério da Saúde: Base de Dados de Síndrome Respiratória Aguda Grave (SRAG). DATASUS, Brasília (2023)
10. Ribeiro, K.B., et al.: Clinical severity of COVID-19 in Brazil: factors associated with ICU admission and mortality. BMC Infectious Diseases 21 (2021)
11. DATASUS, Ministério da Saúde: Banco de Dados de Síndrome Respiratória Aguda Grave (SRAG) 2021–2024. Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2021-a-2024. Acesso em: 09 dez. 2025.
12. Grus, J.: *Data Science do Zero*. Alta Books, Rio de Janeiro (2021). E-book. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788550816463/. Acesso em: 20 nov. 2025.
13. Salles, D. B.: Predição de desfecho desfavorável em pacientes com COVID-19 usando técnicas de aprendizado de máquina. Monografia (Graduação em Ciência da Computação), Universidade Federal de Ouro Preto, Instituto de Ciências Exatas e Biológicas (2023). Disponível em: https://www.monografias.ufop.br/bitstream/35400000/5902/8/MONOGRAFIA$_Predio Pacientes COVID$19.*pdf*.*Acessoem* : 20*nov*.2025.