Tarot da Morte: Predição de Desfecho em Casos de SRAG (2013–2025)

Primeiro Autor: Mayara Vieira Martins Sants[1][0009−0001−7946−7855] e Segundo Autor: Ryan Sousa de Moraes[1][0009−0002−6856−8906]

Instituto Federal de Brasília (IFB), Brasil

Resumo. Este trabalho apresenta um modelo preditivo para desfecho (alta ou óbito) em casos de Síndrome Respiratória Aguda Grave (SRAG) utilizando dados públicos do SIVEP-Gripe (2013–2025). Foram processados 4.173.338 registros com desfecho conhecido, submetidos a etapas de harmonização, pré-processamento e engenharia de features. A variável alvo foi definida como óbito (incluindo outras causas) versus alta. A análise exploratória revelou um desbalanceamento significativo (75,38% altas vs. 24,62% óbitos) e identificou idade avançada, sexo masculino e presença de comorbidades como fatores de risco. Um modelo de regressão logística com ajuste de pesos foi treinado com 10 features clínicas e demográficas. O modelo alcançou AUC-ROC de 0,8262, acurácia de 73,45% e recall de 75% para a classe de óbito, demonstrando capacidade discriminatória moderada-alta. As variáveis mais preditivas foram suporte ventilatório invasivo e internação em UTI, superando fatores basais como idade. O estudo conclui que marcadores de gravidade aguda são mais informativos para prognóstico imediato, oferecendo uma ferramenta viável para vigilância e triagem precoce em saúde pública.

Palavras-Chave: Predição · Aprendizado de Máquina · SRAG · Regressão Logística · Saúde Pública

1 Introdução

O presente trabalho tem como objetivo demonstrar o processo de preparação, análise exploratória e visualização de dados utilizando ferramentas amplamente empregadas em projetos de Ciência de Dados. A partir de um conjunto de dados previamente disponibilizado, foram realizadas etapas como carregamento, inspeção, tratamento e representação gráfica, permitindo observar padrões, tendências e relações entre variáveis de forma clara e objetiva.

No contexto epidemiológico, essas etapas tornam-se especialmente relevantes para o estudo da Síndrome Respiratória Aguda Grave, uma condição monitorada continuamente pelos sistemas oficiais de vigilância em saúde. A disponibilidade de bases históricas, como as fornecidas pelo SIVEP-Gripe, possibilita investigar o comportamento temporal da doença e compreender variações em sua incidência.

Com a aplicação de técnicas de análise e modelagem preditiva, é possível não apenas interpretar dados passados, mas também estimar resultados futuros.

Assim, este trabalho utiliza os dados públicos de SRAG para identificar padrões, visualizar tendências e construir previsões que auxiliem na compreensão da dinâmica da doença. A integração entre análise exploratória e estimativas preditivas contribui para fortalecer o processo de vigilância epidemiológica e oferece suporte ao planejamento de ações em saúde pública.

2 Métodos Utilizados no Projeto

2.1 Coleta e Harmonização de Dados

A base de dados deste estudo foi constituída a partir do Sistema de Informação de Vigilância Epidemiológica da Gripe (**SIVEP-Gripe**), um registro público nacional. Foram coletados todos os registros disponíveis no período de **2013 a 2025**, originalmente distribuídos em arquivos anuais no formato CSV.

Para criar um *dataset* único e analítico, os 13 arquivos CSV foram consolidados em um único banco de dados relacional, utilizando o sistema **SQLite**. O banco resultante, nomeado `srag_mestre_FINAL_V5.db`, serviu como repositório central e estruturado para todas as operações subsequentes.

Em seguida, realizou-se uma etapa crítica de harmonização e padronização. Este processo incluiu a unificação de nomenclaturas de variáveis e categorias entre os diferentes anos. Por exemplo, a condição referida em alguns anos como "METABOLICA" foi padronizada para a denominação "DIABETES". Além disso, foram corrigidas inconsistências técnicas, como problemas de *encoding* de caracteres e variações na estrutura ou nomeação de colunas ao longo da série histórica, garantindo a coerência e a integridade do *dataset* consolidado.

Por fim, aplicou-se um filtro baseado no desfecho clínico. Para as análises e modelagem preditiva que demandavam um *outcome* definido, foram selecionados exclusivamente os registros com desfecho conhecido e documentado, ou seja, aqueles em que a variável `EVOLUCAO` assumiu os valores 1 (Cura), 2 (Óbito) ou 3 (Óbito por outras causas). Esta seleção assegurou que as observações utilizadas nos modelos tivessem a variável *target* válida e não ambígua.

2.2 Pré-processamento e Engenharia de Features

- **Variável alvo (TARGET)**:
  - 0 = Alta (sobrevivência)
  - 1 = Óbito (incluindo óbitos por outras causas)
- **Tratamento de dados**:
  - Conversão de tipos (idade, comorbidades).
  - Binarização de variáveis categóricas (UTI, SUPORT_VEN, comorbidades).
  - Codificação de sexo (CS_SEXO_BIN: M=1, F=0).
  - Tratamento de idades inconsistentes (ex.: >120 anos, valores 9999).
- **Criação de features**:
  - NUM_COMORBIDADES: soma das comorbidades binarizadas.

2.3 Análise Exploratória de Dados (EDA)

Realizou-se uma análise exploratória de dados (EDA) para caracterizar a população do estudo e investigar associações preliminares com o desfecho clínico [4]. A distribuição do desfecho (alta vs. óbito) foi analisada para definir o balanceamento inicial do *dataset*.

A **análise univariada** focou-se nas principais variáveis demográficas: a **idade** foi descrita por estatísticas sumárias (média, mediana, desvio-padrão) e sua distribuição por desfecho foi visualizada via *boxplot*; o **sexo** foi analisado comparando-se a proporção de óbitos entre os gêneros.

A **análise bivariada** examinou relações específicas com a mortalidade. Avaliou-se a associação da internação em Unidade de Terapia Intensiva (UTI) com o desfecho. Em seguida, a influência de comorbidades individuais (ex.: diabetes, obesidade) sobre a taxa de óbito foi verificada. Adicionalmente, testou-se a hipótese de uma relação dose-resposta, analisando se um maior número de comorbidades por paciente estava associado a uma taxa de óbito crescente.

Todas as análises foram suportadas por visualizações apropriadas, incluindo gráficos de barras, *boxplots*, histogramas e *heatmaps* para correlações, geradas com as bibliotecas *matplotlib* e *seaborn* do Python.

2.4 Modelagem Preditiva

Para a predição do desfecho clínico, o algoritmo de Regressão Logística foi selecionado, implementado através da função `LogisticRegression` da biblioteca SCIKIT-LEARN. A escolha deste modelo fundamenta-se em sua adequação para problemas de classificação binária, sua alta interpretabilidade—permitindo a análise do peso e direção de cada *feature*—e seu amplo uso e validação em estudos clínicos e epidemiológicos [2].

Considerando a distribuição desbalanceada do *dataset* (aproximadamente 75% de altas versus 25% de óbitos), o modelo foi configurado com o parâmetro `class_weight='balanced'`. Esta estratégia atribui pesos inversamente proporcionais à frequência de cada classe durante o treinamento, corrigindo o viés do modelo em favorecer a classe majoritária.

Os dados foram divididos em conjuntos para treinamento e validação, adotando-se uma proporção de 80% para treino e 20% para teste. A divisão foi estratificada com base na variável *target*, garantindo que a proporção de desfechos (alta/óbito) se mantivesse representativa em ambas as partições.

O modelo final foi treinado utilizando um conjunto de 10 *features* clínicas e demográficas previamente selecionadas e tratadas: `NU_IDADE_N` (idade normalizada), `CS_SEXO_BIN` (sexo binarizado), `UTI_BIN` (internação em UTI), `SUPORT_VEN_BIN` (suporte ventilatório), `DIABETES_BIN`, `CARDIOPATI_BIN`, `ASMA_BIN`, `RENAL_BIN`, `PNEUMOPATI_BIN` (comorbidades binarizadas) e `NUM_COMORBIDADES` (contagem total de comorbidades).

2.5 Avaliação do Modelo

A avaliação do desempenho do modelo preditivo foi conduzida com base em um conjunto abrangente de métricas, calculadas a partir do conjunto de teste. Foram reportadas a acurácia global, bem como as métricas específicas por classe: precisão, *recall* e *F1-Score*. A Matriz de Confusão foi analisada para detalhar a natureza dos acertos e erros, discriminando falsos positivos e falsos negativos.

A métrica principal adotada para avaliar a capacidade discriminativa do modelo foi a Área Sob a Curva da Característica de Operação do Receptor (AUC-ROC). Esta medida, que avalia o desempenho do classificador em todos os limiares de decisão possíveis, é particularmente robusta em cenários com desbalanceamento de classes [1].

Complementarmente à avaliação de desempenho, realizou-se uma interpretação dos coeficientes do modelo de regressão logística. Os coeficientes padronizados foram analisados para inferir a importância relativa e a direção da associação de cada *feature* com o desfecho de óbito. Esta análise permite uma compreensão explanatória do modelo, identificando quais variáveis apresentam maior peso preditivo no contexto clínico estudado.

2.6 Ferramentas e Bibliotecas

Para a implementação deste trabalho, foi adotada a linguagem de programação Python, na versão 3.8. O ambiente de execução escolhido foi o Google Colab, que fornece uma plataforma computacional em nuvem. As operações de manipulação e análise de dados foram realizadas com o auxílio das bibliotecas Pandas e NumPy. Para o armazenamento persistente e consulta estruturada dos dados consolidados, utilizou-se o sistema de banco de dados SQLite, por meio do módulo SQLite3. A geração de visualizações e gráficos analíticos foi conduzida com as bibliotecas Matplotlib e Seaborn, e por fim, as etapas de modelagem preditiva e avaliação de desempenho dos algoritmos foram implementadas utilizando a biblioteca Scikit-learn.

2.7 Fluxo Metodológico

O fluxo metodológico desenvolvido para este estudo foi estruturado em etapas sequenciais e interligadas, conforme detalhado a seguir:

1. **Extração e Armazenamento de Dados:** Os dados brutos, provenientes do SIVEP-Gripe, foram coletados e armazenados de forma estruturada em um banco de dados relacional SQLite. Esta abordagem assegurou a integridade dos dados e facilitou consultas eficientes durante as fases subsequentes.
2. **Limpeza e Pré-processamento:** Na etapa de limpeza, realizou-se a definição da variável **target** (alvo) do modelo. Para este problema de classificação binária, a variável foi codificada como 0 (Alta) e 1 (Óbito). Adicionalmente, a variável numérica 'idade' foi submetida a tratamento específico para valores inconsistentes (ex.: >120 anos, valores 9999), a fim de adequá-la à modelagem.

3. **Análise Exploratória de Dados (EDA):** Conduziu-se uma análise exploratória abrangente, contemplando tanto uma **análise descritiva** (estatísticas de tendência central e dispersão) quanto uma **análise visual**. Esta etapa teve por objetivo compreender a distribuição das variáveis, identificar relações preliminares e validar a qualidade dos dados após o pré-processamento.
4. **Engenharia de Features (Feature Engineering):** Para preparar os atributos para os algoritmos de aprendizado de máquina, aplicaram-se técnicas de transformação. Isto incluiu a **binarização** de variáveis categóricas (UTI, SUPORT_VEN, comorbidades), a **codificação** de sexo (CS_SEXO_BIN: M=1, F=0), e a criação de novas *features* derivadas, como a variável de **contagem** NUM_COMORBIDADES.
5. **Modelagem Preditiva:** O núcleo da predição foi implementado por meio de um modelo de **Regressão Logística**. Considerando o desequilíbrio na distribuição das classes da variável target (75% alta vs. 25% óbito), a técnica de **balanceamento de dados** via ajuste de pesos (`class_weight='balanced'`) foi empregada para mitigar vieses e melhorar a capacidade de generalização do modelo.
6. **Avaliação de Desempenho:** O modelo final foi avaliado utilizando um conjunto de métricas robustas e complementares. A métrica principal foi a **Área Sob a Curva ROC (AUC-ROC)**, que avalia o desempenho geral em termos de discriminação entre classes. A análise foi complementada pela inspeção da **Matriz de Confusão**, que detalha acertos e erros por classe, e por um **Relatório de Classificação** abrangente, apresentando métricas como precisão (*precision*), revocação (*recall*) e medida F1 (*F1-score*) para cada classe.

3 Resultados

3.1 Estatísticas Descritivas e Análise Exploratória

O conjunto final utilizado para análise contou com 4.173.338 registros de casos de SRAG com desfecho conhecido, no período de 2013 a 2025. A Tabela 1 resume a distribuição da variável alvo e das principais variáveis demográficas e clínicas.

A análise de desbalanceamento revelou uma distribuição significativamente desigual entre as classes (Figura 1). A idade demonstrou ser um forte discriminador, com pacientes que evoluíram a óbito apresentando média etária 23 anos superior aos sobreviventes.

3.2 Análise Bivariada e Comorbidades

A Tabela 2 apresenta as taxas de óbito estratificadas por características clínicas e comorbidades.

A internação em UTI foi o fator com maior associação bruta com óbito, aumentando o risco em mais de 3 vezes. A análise de contagem de comorbidades revelou um padrão de dose-resposta, onde pacientes com 3 ou mais comorbidades apresentaram taxas de óbito superiores a 40% (Figura 2).

Table 1. Estatísticas descritivas da amostra final (n = 4.173.338).

Variável	**Categoria/Medida**	**Amostra (%)**
Desfecho (TARGET)	Alta (0)	3.144.326 (75,38%)
	Óbito (1)	1.029.012 (24,62%)
Idade (NU_IDADE_N)	Média (DP) – Total	49,3 (28,1) anos
	Média – Alta	43,5 (28,3) anos
	Média – Óbito	66,5 (18,1) anos
Sexo (CS_SEXO)	Masculino	2.221.122 (53,21%)
	Feminino	1.952.216 (46,79%)
UTI	Internados	1.229.674 (29,47%)
	Não internados	2.943.664 (70,53%)

Table 2. Taxa de óbito por variáveis clínicas selecionadas.

Variável	**Taxa de Óbito (%)**
Sexo: Masculino	25,34
Sexo: Feminino	23,79
UTI: Sim	43,75
UTI: Não	13,84
Cardiopatia	31,16
Doença Renal	39,05
Diabetes	18,45
Asma	31,16
Pneumopatia	42,69

3.3 Desempenho do Modelo Preditivo

O modelo de Regressão Logística com ajuste de peso (`class_weight='balanced'`) foi treinado com 3.338.138 amostras e testado com 834.535 amostras, mantendo a mesma proporção de óbitos (24,62%) em ambos os conjuntos.

As métricas de desempenho estão resumidas na Tabela 3.

A Matriz de Confusão (Tabela 4) detalha o desempenho por classe:

3.4 Importância das Variáveis no Modelo Final

A interpretação dos coeficientes padronizados do modelo de regressão logística final (Figura 3) permitiu identificar as variáveis com maior poder preditivo para o desfecho de óbito, quantificando sua magnitude e direção de associação.

As *features* que demonstraram maior associação positiva com a mortalidade foram: **suporte ventilatório invasivo** (SUPORT_VEN_BIN; Coef. = +1,784) e **internação em Unidade de Terapia Intensiva** (UTI_BIN; Coef. = +0,841). Em contrapartida, a presença de **asma** (ASMA_BIN; Coef. = -0,466) apresentou uma associação negativa com o óbito no modelo ajustado. Outras

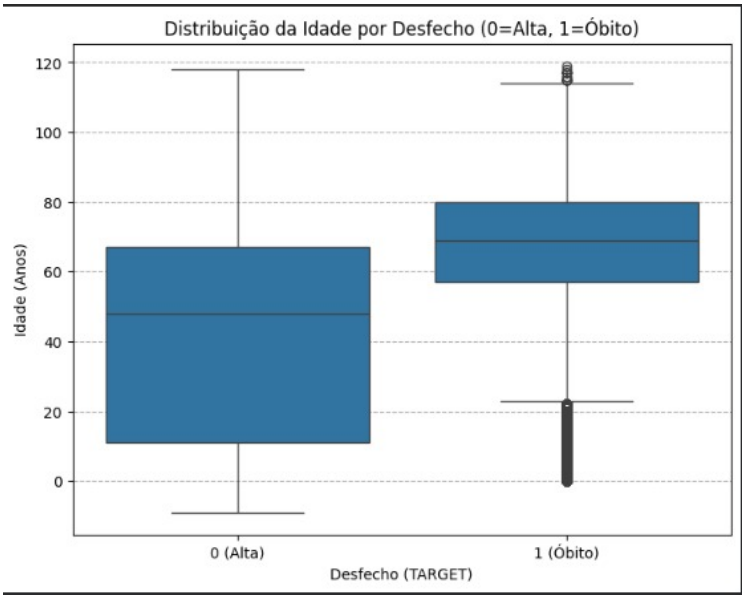

Fig. 1. Distribuição percentual da variável alvo TARGET (0 = Alta, 1 = Óbito).

Table 3. Métricas de avaliação do modelo preditivo.

Métrica	**Valor**
Acurácia	73,45%
AUC-ROC	0,8262
Precisão (Óbito)	0,47
Recall/Sensibilidade (Óbito)	0,75
F1-Score (Óbito)	0,58

comorbidades com influência preditiva positiva foram a **doença renal crônica** (RENAL_BIN; Coef. = +0,406) e o **sexo masculino** (CS_SEXO_BIN; Coef. = +0,173).

Um resultado digno de nota refere-se à variável **idade** (NU_IDADE_N), que apresentou um coeficiente relativamente baixo (+0,042) no modelo multivariado. Esta observação sugere que o efeito bruto da idade sobre a mortalidade pode ser, em parte, mediado ou confundido por variáveis de gravidade clínica aguda, como a necessidade de UTI e de suporte ventilatório, as quais capturam uma parcela significativa do risco associado aos anos de vida [3].

4 Considerações Finais

Este estudo demonstrou a viabilidade metodológica de construir um modelo preditivo para o desfecho clínico em SRAG, a partir da integração e análise de

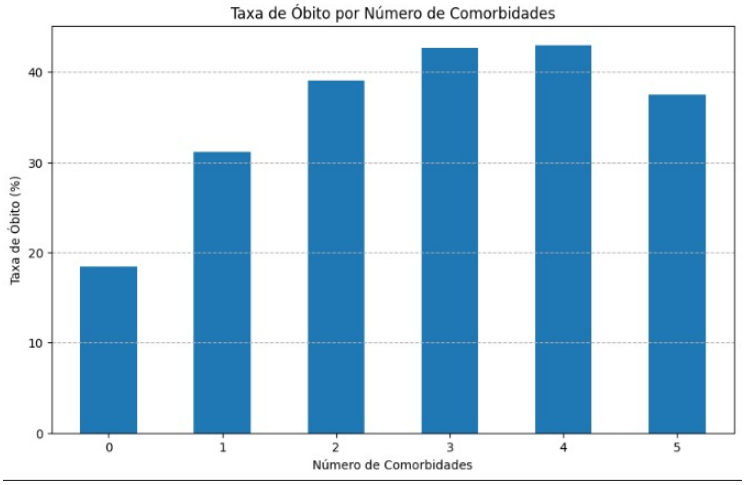

Fig. 2. Taxa de óbito (%) por número de comorbidades presentes (0 a 5).

Table 4. Matriz de confusão no conjunto de teste (n = 834.535).

	Predito: Alta	**Predito: Óbito**
Real: Alta	459.677 (VN)	169.411 (FP)
Real: Óbito	52.191 (FN)	153.256 (VP)

dados públicos de larga escala do SIVEP-Gripe. O processo envolveu um *pipeline* completo, desde a consolidação e harmonização de bases históricas (2013-2025) até a aplicação e avaliação de técnicas de aprendizado de máquina supervisionado para classificação binária.

Os principais achados podem ser assim sintetizados:

Padrões Epidemiológicos Consolidados: Os resultados confirmaram, em uma série temporal extensa, que idade avançada, sexo masculino e a presença de comorbidades prévias atuam como fatores de risco independentes para óbito por SRAG, com destaque para doença renal crônica e cardiopatia.

Marcadores de Gravidade como Preditores Primários: Um dos achados mais robustos foi a superior capacidade preditiva das variáveis de intervenção clínica (internação em UTI e necessidade de suporte ventilatório invasivo) sobre fatores demográficos e comorbidades basais. Este resultado sugere que os marcadores da gravidade da doença aguda são mais informativos para o prognóstico imediato do que as condições de saúde prévias do paciente.

Modelo com Bom Desempenho Discriminatório: O modelo de Regressão Logística balanceada alcançou uma AUC-ROC de 0,826, indicando uma capacidade discriminativa moderada-alta para diferenciar entre desfechos de alta e óbito. Este nível de desempenho é adequado para aplicações em ferramentas de apoio à triagem e vigilância em saúde pública.

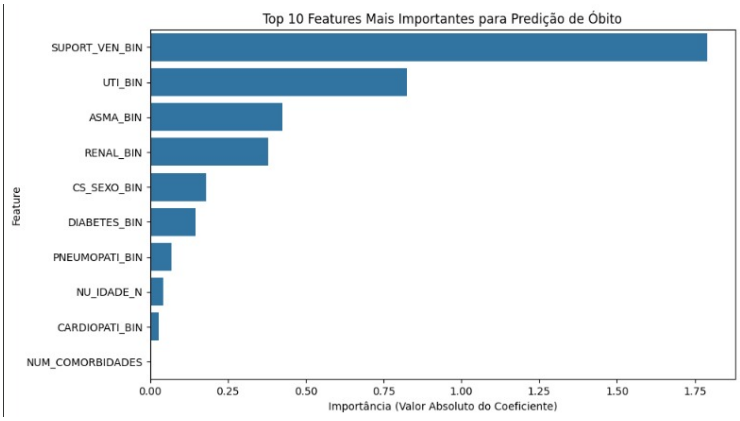

Fig. 3. Importância relativa das features (valor absoluto dos coeficientes).

Compromisso entre Sensibilidade e Precisão: A estratégia de balanceamento de classes permitiu uma sensibilidade (*recall*) elevada para a classe de óbito, prioritária para um cenário clínico de alerta precoce. Esse ganho, no entanto, ocorreu às custas de uma menor precisão, refletindo o compromisso inerente à seleção do ponto de corte operacional, apropriado para contextos onde a identificação de casos graves é fundamental.

Limitações e Aprendizados Técnicos: O processo extensivo de harmonização de dados revelou desafios significativos na integração de bases históricas, incluindo mudanças na nomenclatura de variáveis e inconsistências de codificação. Esta experiência reforça a importância crítica da manutenção de metadados robustos e documentação contínua em sistemas nacionais de vigilância epidemiológica.

4.1 Implicações Práticas

O modelo desenvolvido pode ser incorporado a sistemas de vigilância em saúde como ferramenta de alerta precoce para identificar pacientes com maior risco de evolução desfavorável. Sua implementação permitiria otimizar alocação de recursos em UTIs e antecipar necessidades de suporte ventilatório.

4.2 Limitações

As principais limitações incluem: (1) viés de notificação inerente a sistemas de vigilância; (2) ausência de variáveis laboratoriais detalhadas no modelo; (3) não consideração de variantes virais e status vacinal; (4) desempenho subótimo em precisão para a classe de óbito.

4.3 Contribuições

Este estudo oferece contribuições metodológicas e empíricas para o campo da análise de dados em saúde. Em termos metodológicos, propõe e documenta um *pipeline* completo de dados, reprodutível e de código aberto, que abrange desde a extração e harmonização de registros administrativos complexos até a modelagem preditiva avançada, servindo como um *template* para estudos similares.

Empiricamente, o trabalho gera evidências robustas sobre os principais preditores de óbito por SRAG, derivadas da análise de uma série temporal extensa de 13 anos. Os resultados destacam o papel preponderante da gravidade clínica aguda (suporte ventilatório e internação em UTI) sobre fatores demográficos tradicionais em modelos multivariados.

Adicionalmente, a pesquisa contribui com a documentação detalhada dos desafios práticos inerentes à harmonização de dados de vigilância epidemiológica de longo prazo, um conhecimento tácito valioso para a comunidade. Por fim, o projeto resultou na consolidação do modelo preditivo final como uma *baseline* pública, disponível para validação externa e aprimoramentos futuros por outros pesquisadores, e cumpriu seu objetivo didático ao proporcionar o aprendizado aplicado e a formação em pesquisa para dois estudantes de tecnologia.

References

1. FAWCETT, Tom. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861–874, 2006.
2. HOSMER Jr, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3rd ed. John Wiley & Sons, 2013.
3. KLEINBERG, J.; MULLAINATHAN, S. The algorithmic bottleneck: How machine learning shapes inequality. *Science*, v. 361, n. 6401, p. 749–750, 2018.
4. WILKINSON, Leland. **The Grammar of Graphics**. 2nd ed. New York: Springer, 2005.
5. Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
6. Harris, C.R., et al. Array programming with NumPy. *Nature*, vol. 585, pp. 357–362, 2020.
7. McKinney, W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, pp. 56–61.
8. Waskom, M.L. Seaborn: statistical data visualization. *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.
9. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.
10. Ministério da Saúde (Brasil). Banco de Dados de Síndrome Respiratória Aguda Grave (SRAG). SIVEP-Gripe, 2025. Disponível em: `https://sivepgripe.saude.gov.br/`