Modelagem Preditiva: Fatores Demográficos e Regionais Determinantes para Cuidados Intensivos em Casos de SRAG no Brasil

Davi Henrique Menezes da Cruz [1][0000-1111-2222-3333] e Sabrina de Oliveira Souza[2][1111-2222-3333-4444]

[1] Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB), Brasil
davih1662@gmail.com
[2] Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB), Brasil
so38196@gmail.com

Resumo. Este estudo analisa a aplicabilidade da Regressão Logística na predição de internações em UTI para pacientes com Síndrome Respiratória Aguda Grave (SRAG) por meio dos dados obtidos pelo OpenDataSUS no período de 2016 a 2018. Avaliou-se o impacto de variáveis demográficas e regionais no desfecho clínico. A metodologia incluiu limpeza de dados, codificação *one-hot* e normalização. Inicialmente, o modelo obteve AUC-ROC de 0,69 e Recall de 27%. A otimização pelo ajuste do limiar para 0,3 elevou o Recall para 76,57%. Embora a precisão tenha caído para 45,28%, o *trade-off* é justificado pelo fato de que em cenários de risco, priorizar a sensibilidade garante a identificação de casos graves e a correta alocação de recursos hospitalares.

Palavras-chave: SRAG, Regressão Logística, Perfil Demográfico.

1 Introdução

A Síndrome Respiratória Aguda Grave (SRAG) é um importante problema de saúde pública no Brasil. Trata-se de "uma complicação que pode surgir a partir de diversas infecções respiratórias" (Pinto, 2025). A identificação precoce de pacientes com risco de internação em Unidade de Terapia Intensiva (UTI) é fundamental para a otimização do atendimento e a melhoria dos desfechos clínicos.

Este estudo analisa os fatores associados à internação em UTI, considerando variáveis demográficas, clínicas e de acesso à saúde. O foco é desenvolver um modelo preditivo capaz de identificar a probabilidade de cuidados intensivos utilizando dados de 2016 a 2018.

A escolha desses anos tomados como as bases de dados permitiu uma visão abrangente dos padrões da SRAG antes da pandemia de COVID-19, oferecendo uma linha de base para estudos comparativos. Dessa forma, o modelo atua como uma ferramenta estratégica para o planejamento de recursos hospitalares e para o suporte à decisão clínica em cenários de alta demanda.

2 Materiais e Métodos

As três bases de dados referentes aos anos de 2016, 2017 e 2018 foram unificadas, resultando em um conjunto inicial de 131.687 registros e 112 colunas. Em seguida, foi realizada a seleção de variáveis, reduzindo o escopo para 21 colunas relevantes, e iniciada a etapa de limpeza e preparação dos dados. Esta fase incluiu o tratamento de valores ausentes (NaN) e inconsistentes (código 9 para "Ignorado"), com destaque para a variável raça/cor, que apresentava o maior índice de ruído, com 13,06% de registros ignorados e 3,10% de valores ausentes.

Foram realizadas padronizações essenciais: a variável *data_primeiros_sintomas* foi convertida para o formato *datetime64[ns]*, e a variável *idade* foi normalizada para garantir a estabilidade estatística do modelo. Para assegurar a integridade da análise de desfecho, foram removidos todos os registros com informações ausentes ou ignoradas nas colunas críticas de *uti* e *evolucao.* Após estes ajustes, foi obtido um conjunto final de 117.403 registros válidos.

Objetivando fazer uma modelagem de qualidade, foi criada a variável alvo binária *UTI_BINARIO*. Casos com indicação positiva de UTI foram classificados como 1 significando “Sim”, totalizando 41.142 registros, enquanto os demais foram classificados como 0 para “Não", totalizando 76.261 registros. As variáveis categóricas, como sexo, raça/cor e unidade federativa (UF), foram transformadas em variáveis *dummy* por meio de one-hot encoding, resultando em um conjunto final de 62 atributos preditores, ou seja, as features.

O modelo de aprendizado de máquina selecionado para este estudo foi a Regressão Logística. Segundo Grus (2021), esse modelo é uma das ferramentas fundamentais para problemas de classificação binária devido à sua simplicidade e eficácia. Esse modelo permite a análise clara dos coeficientes de impacto de cada fator demográfico e clínico, possibilitando a extração da Razão de Chances. Essa característica é essencial no contexto da saúde pública, pois permite identificar não só a predição de risco, mas quais variáveis são mais determinantes para o desfecho clínico.

Além da eficiência computacional para lidar com o volume de 117.403 registros, o modelo oferece uma vantagem estratégica central: a geração de saídas em termos de probabilidade através da função sigmóide. Essa flexibilidade matemática foi o que permitiu a otimização do projeto por meio do ajuste do limiar de decisão. Ao manipular o ponto de corte, foi possível transformar um modelo inicialmente conservador em uma ferramenta de triagem sensível, priorizando o Recall para garantir a segurança clínica.

O rigor metodológico foi assegurado pela divisão do conjunto de dados em treino e teste na proporção de 80% e 20%, respectivamente, resultando em 93.922 registros para o treinamento e 23.481 para o teste. A divisão foi realizada com amostragem estratificada, garantindo a manutenção da proporção da variável alvo em ambos os conjuntos e assegurando a validade estatística da avaliação final, que obteve

um AUC-ROC inicial de 0,6925. Esse valor confirma que, apesar da simplicidade linear do modelo, ele possui uma capacidade de distinção razoável entre os grupos de UTI, servindo como uma linha de base sólida para as análises de risco propostas.

3 Resultados

O desempenho inicial do modelo, fazendo uso do limiar de decisão padrão de 0,5, apresentou uma Acurácia de 67,91% e um AUC-ROC de 0,6925. Embora a acurácia mostre um resultado satisfatório, a análise detalhada das métricas identificou uma fragilidade crítica: o Recall foi de apenas 27,00%.

A Matriz de Confusão (Fig. 1) inicialmente evidenciou esse problema: o modelo classificou corretamente 13.725 pacientes que não precisam de UTI (Verdadeiros Negativos), mas falhou em identificar 6.007 casos graves, rotulando-os como Falsos Negativos. Na prática, o modelo ignorou 73% dos pacientes em risco, um erro grave, pois em um contexto hospitalar onde a omissão de socorro intensivo pode ser fatal. A Curva ROC, com área de 0,6925, confirmou que o modelo possuía uma capacidade de distinção razoável, mas que o ponto de corte padrão estava muito conservador.

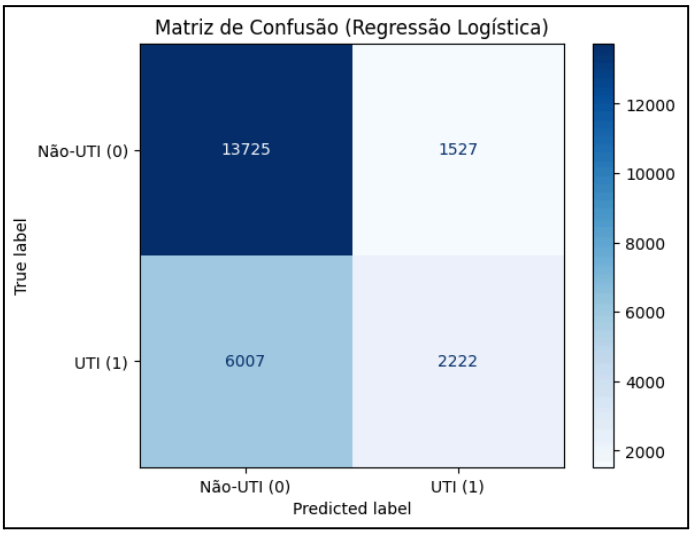

Fig. 1. Matriz de confusão

Diante da necessidade de reduzir os Falsos Negativos, optou-se por realizar um teste de limiar. Ao reduzir o ponto de corte de 0,5 para 0,3, observou-se uma transformação significativa na performance:

- Recall (Sensibilidade): Subiu para 76,57%.
- Precisão: Recuou para 45,28%.

- Acurácia: Ajustou para 59,35%.

Esse ajuste permitiu que o modelo capturasse a maioria dos pacientes graves (76,57%), cumprindo sua função primordial de triagem. O *trade-off* observado na precisão indica que 54,72% dos alertas de UTI serão Falsos Positivos. Porém, na gestão de saúde pública, o custo de um alarme falso, que impacta da sobrecarga logística, é secundário ao custo de um falso negativo, que significa um risco de vida.

Buscando alternativas ao ajuste de limiar, aplicou-se a técnica SMOTE (Synthetic Minority Over-sampling Technique) para balancear as classes de treino, totalizando 61.009 registros para cada classe. Embora essa técnica tenha melhorado ligeiramente a Precisão para 0,4971 em comparação ao limiar 0,3, ele entregou um Recall de apenas 0,5588. Como o objetivo central é a segurança do paciente e o máximo de sensibilidade, o modelo SMOTE foi rejeitado em favor do modelo de Regressão Logística com limiar de 0,3.

A análise exploratória de dados (EDA) (Fig. 2) forneceu o contexto visual para as predições do modelo. O gráfico de Proporção de Internação por Raça/Cor revelou disparidades importantes, enquanto a Contagem de Casos por Sexo (Fig. 3) mostrou um equilíbrio volumétrico, mas com nuances nas taxas de gravidade.

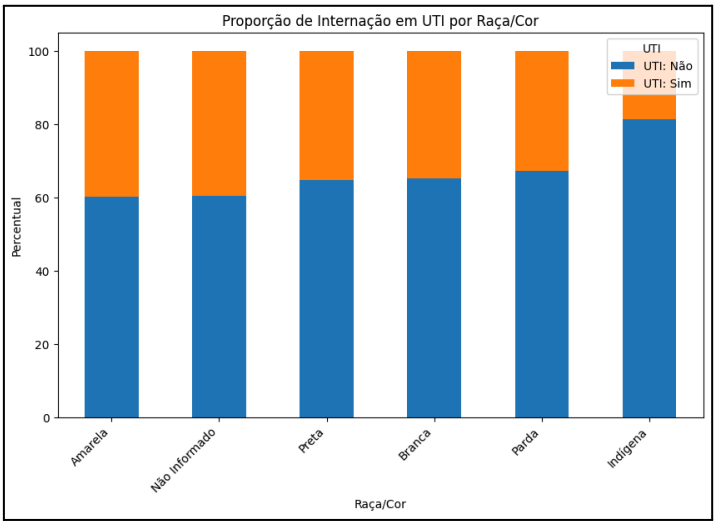

Fig. 2. Proporção de internação em UTI por raça/cor

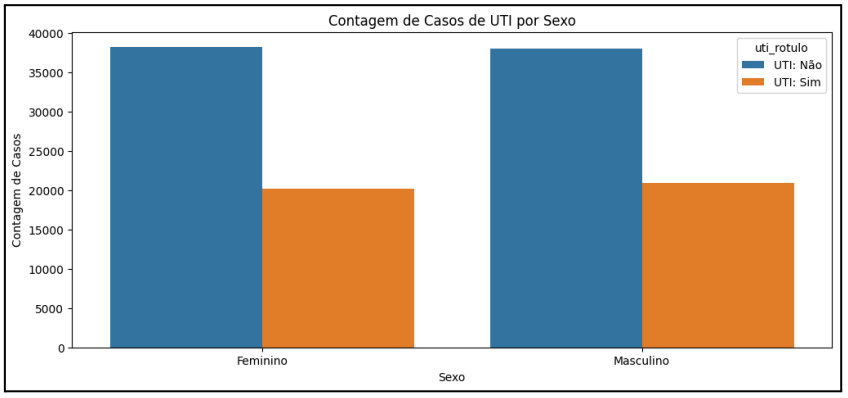

Fig. 3. Contagem de Casos por Sexo

O Boxplot de Distribuição da Idade (Fig. 4) demonstrou que, embora a média de idade seja similar entre os grupos, a dispersão e a presença de *outliers* na base de dados influenciam a probabilidade de evolução para a UTI. Por fim, a análise das Unidades Federativas (UF) (Fig. 5) destacou variações regionais acentuadas na ocupação de leitos, sugerindo que o CEP do paciente é um forte componente preditivo da disponibilidade e necessidade de tratamento intensivo no Brasil.

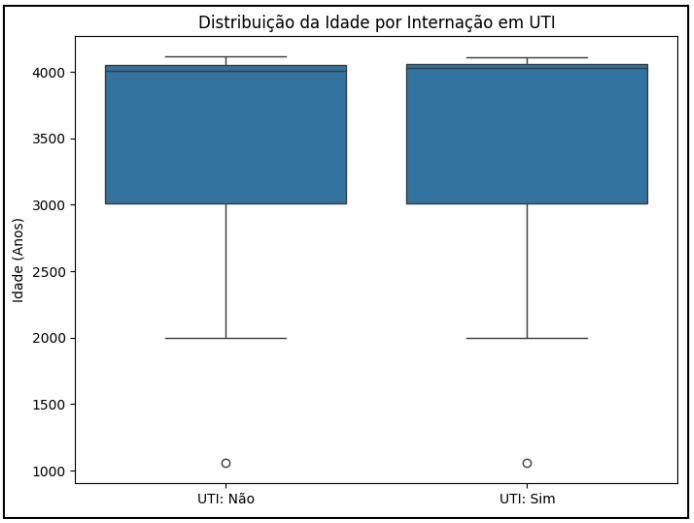

Fig. 4. Boxplot de Distribuição da Idade

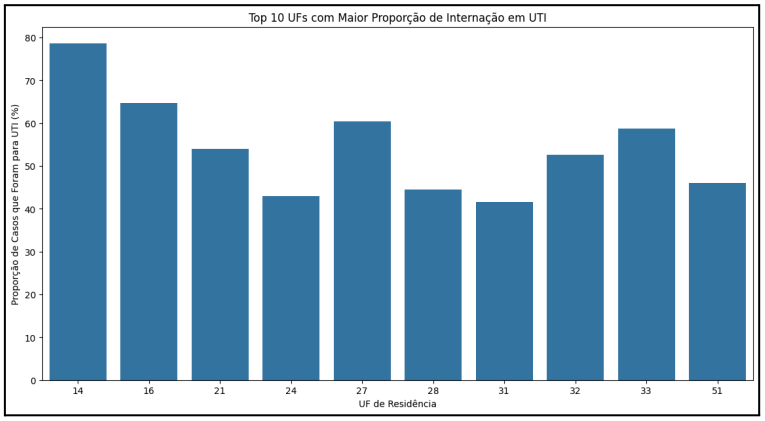

Fig. 5. Unidades Federativas com maior proporção de internação

3 Considerações finais

O presente estudo demonstrou que a aplicação da Regressão Logística, constitui uma ferramenta poderosa para a triagem e predição de gravidade em pacientes com SRAG no Brasil. A análise dos dados revelou que o perfil demográfico, regional e clínico do paciente possui um valor preditivo significativo na determinação da necessidade de suporte em Unidade de Terapia Intensiva (UTI).

A principal contribuição técnica deste trabalho foi a demonstração de que métricas de desempenho puramente matemáticas, como a acurácia global, podem ser enganosas em contextos de saúde pública com dados desbalanceados. O modelo inicial, embora apresentasse uma acurácia de 67,91%, falhava em identificar 73% dos casos graves. A decisão metodológica de priorizar o Recall através do ajuste do limiar de decisão para 0,3 provou ser a estratégia mais eficaz, elevando a capacidade de identificação de pacientes críticos para 76,57%.

Embora esse ajuste implique um aumento de falsos positivos, a fundamentação da pesquisa sustenta que, em cenários de risco de vida, a minimização dos falsos negativos é imperativa para evitar a omissão de socorro. A comparação com técnicas de sobreamostragem como o SMOTE reforçou essa escolha, uma vez que o ajuste de limiar ofereceu uma sensibilidade superior para a proteção do paciente.

Referencias

1. GRUS, Joel. **Data Science do Zero: primeiras regras com o Python**. Tradução de Welington Nascimento. 2. ed. Rio de Janeiro: Alta Books, 2021.
2. **OpenDataSUS.** SRAG 2013 a 2018 - Banco de Dados de Síndrome Respiratória Aguda Grave. Brasília, DF: Ministério da Saúde, (s.d.). Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2013-2018. Acesso em: 17 dez. 2025.
3. Pinto, Maria Isabel de Moraes. **Síndrome respiratória aguda grave (SRAG) em alta acende alerta no país.** Nav, 2025. Disponível em: https://nav.dasa.com.br/blog/sindrome-respiratoria-aguda-grave. Acesso em: 17 dez. 2025.