Predição de Óbito e Fatores de Risco em Pacientes com SRAG: Uma Abordagem com Regressão Logística em Dados do OpenDataSUS (2013-2018)

Virginia Prettz Camara Canto[1], Pedro Medeiros[1], Vinicius Henrique Barbosa de Oliveira[1]

[1]Curso de Sistemas para Internet – [Instituto Federal de Brasília] [Brasília] – [DF] – Brasil

{virginia60882@estudante.ifb.edu.br,pedro60883@estudante.ifb.edu.br,vinicius.oliveira1@estudante.ifb.edu.br}

Abstract. *This paper presents a predictive analysis of mortality in patients with Severe Acute Respiratory Syndrome (SARS), using OpenDataSUS data from 2013 to 2018. A machine learning pipeline based on Logistic Regression was developed to classify cases into Cure or Death. Data preprocessing included cleaning, imputation, and binarization. The results showed an AUC-ROC of 0.79. Adjusting the decision threshold to 0.60 improved accuracy to 78% and precision to 35%, reducing false positives, which demonstrates the model's potential for clinical triage.*

Resumo. Este artigo apresenta uma análise preditiva de óbito em pacientes com Síndrome Respiratória Aguda Grave (SRAG), utilizando dados do OpenDataSUS de 2013 a 2018. Desenvolveu-se um pipeline de aprendizado de máquina com Regressão Logística para classificar desfechos em Cura ou Óbito. O pré-processamento incluiu limpeza, imputação e binarização. Os resultados apontaram uma AUC-ROC de 0.79. O ajuste do limiar de decisão para 0.60 elevou a acurácia para 78% e a precisão para 35%, reduzindo falsos positivos e demonstrando potencial para triagem clínica.

1. Introdução

A Síndrome Respiratória Aguda Grave (SRAG) representa um dos desafios mais persistentes para o sistema de saúde pública no Brasil. Caracterizada por quadros respiratórios severos que frequentemente exigem hospitalização e cuidados intensivos, a SRAG impõe uma pressão significativa sobre a infraestrutura hospitalar, especialmente em períodos de surtos sazonais ou pandêmicos. Nesse cenário, a capacidade de identificar precocemente os pacientes com maior risco de evolução para óbito torna-se crucial para a gestão clínica e a alocação eficiente de recursos escassos, como leitos de UTI e suporte ventilatório.

Nos últimos anos, a digitalização dos registros de saúde e a política de dados abertos governamentais, exemplificada pelo portal OpenDataSUS, criaram oportunidades inéditas para a aplicação de técnicas de Ciência de Dados na vigilância epidemiológica. O Sistema de Informação de Vigilância Epidemiológica da Gripe (SIVEP-Gripe) acumula um vasto histórico de notificações que, se devidamente processados e analisados, podem revelar padrões ocultos sobre a gravidade da doença que escapam à análise estatística tradicional.

No entanto, o uso desses dados brutos apresenta desafios técnicos significativos, incluindo a alta dimensionalidade, a presença de valores ausentes e o desbalanceamento

entre as classes de desfecho (cura e óbito). Superar essas barreiras para construir ferramentas preditivas confiáveis é uma etapa fundamental para transformar dados históricos em inteligência acionável.

O objetivo deste trabalho é desenvolver e avaliar um modelo de Aprendizado de Máquina supervisionado capaz de prever o desfecho clínico (Cura ou Óbito) de pacientes notificados com SRAG. Utilizando uma base de dados abrangente do período de 2013 a 2018, o estudo investiga a relação entre variáveis demográficas, sintomas e comorbidades com a mortalidade. Além da predição, o trabalho foca na interpretabilidade dos fatores de risco — com destaque para a influência da idade — e na análise do impacto do ajuste de limiares de decisão na sensibilidade e precisão do modelo, visando sua aplicabilidade em cenários de triagem hospitalar.

2. Materiais e Métodos

A metodologia seguiu o processo de descoberta de conhecimento em bases de dados (KDD), utilizando a linguagem Python e bibliotecas de análise de dados.

2.1. Coleta e Pré-processamento

Os dados foram extraídos dos arquivos anuais do OpenDataSUS (2013 a 2018). A base bruta foi unificada e submetida à seleção de variáveis, mantendo-se atributos demográficos (Idade, Sexo), sintomas (ex: Febre, Tosse, Dispneia) e comorbidades (ex: Cardiopatia, Diabetes).

No tratamento da qualidade dos dados, a coluna diarreia foi removida por conter 100% de valores nulos. A idade foi convertida para formato numérico e valores ausentes foram imputados pela mediana. O alvo evulução foi filtrado para considerar apenas Cura (0) e Óbito (1), resultando em 29.792 registros válidos. Variáveis categóricas passaram por codificação *One-Hot* e sintomas foram binarizados.

3. Resultados

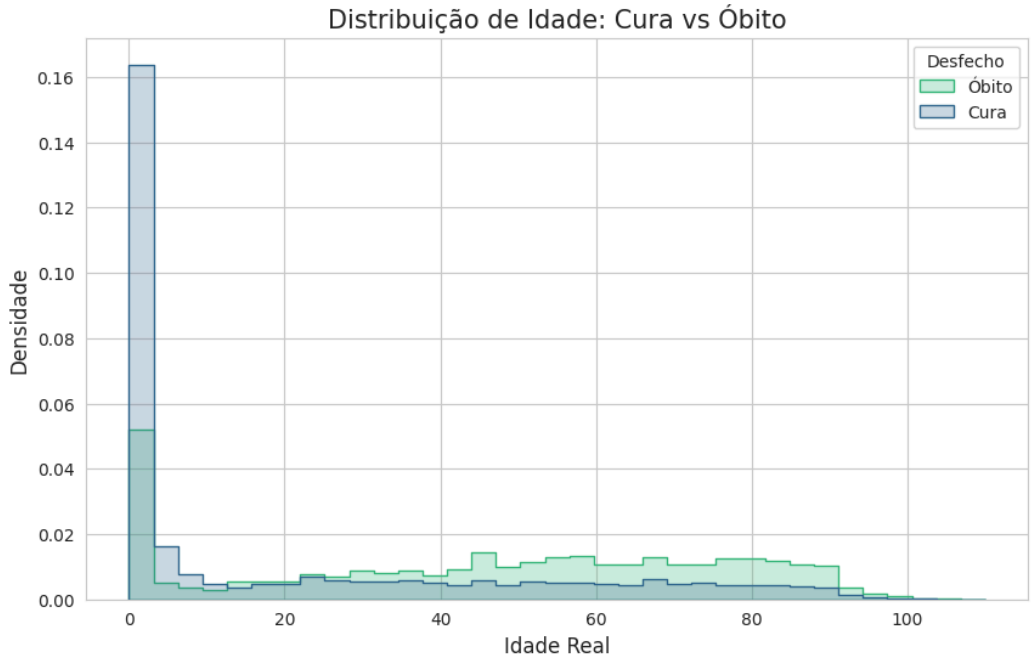
A análise exploratória inicial confirmou que a idade avançada é um fator de risco relevante. A distribuição de idade separada pelo desfecho (Cura vs. Óbito) evidenciou essa correlação.

3.1. Desempenho do Modelo (V1) A primeira versão do modelo (limiar 0,50) obteve uma **AUC-ROC de 0,79** e uma acurácia geral de 73%. A matriz de confusão demonstrou uma alta sensibilidade (*Recall* de 72% para óbito), indicando eficácia na detecção de casos graves, porém com baixa precisão (31%), gerando muitos falsos positivos.

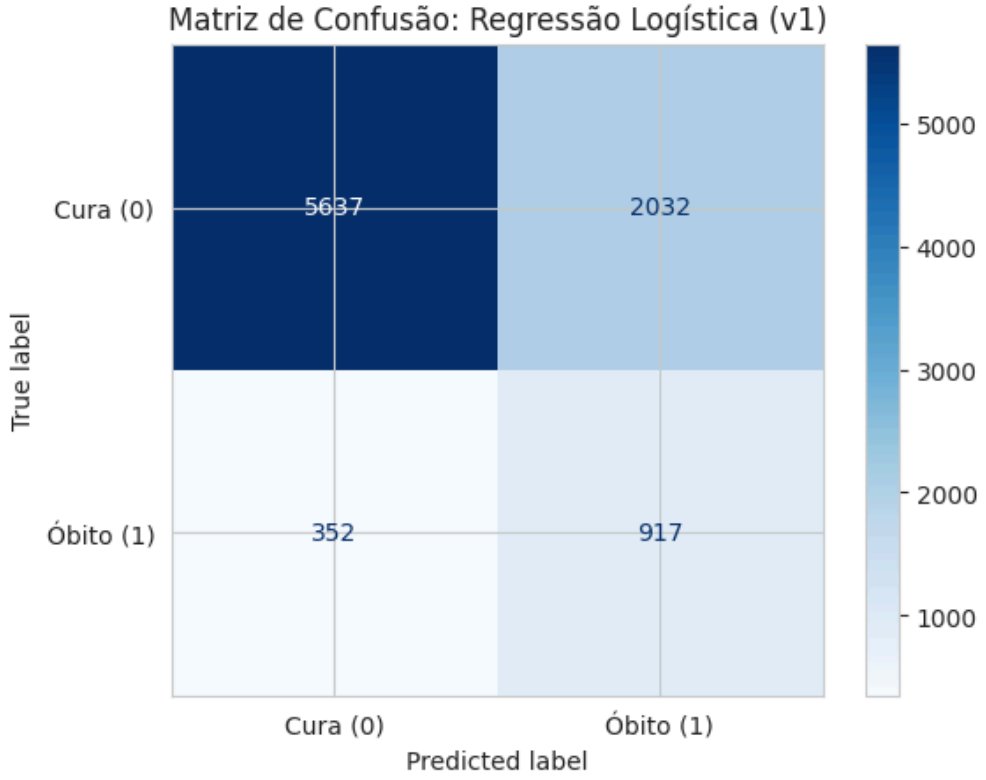
3.2. Otimização do Limiar (V2) Para reduzir a fadiga de alarmes falsos, o limiar de decisão foi ajustado para 0,60. Comparado ao modelo V1, esta alteração resultou em uma redução de falsos positivos de 2.032 para 1.449 e um aumento na acurácia para 78%. Houve um aumento na precisão (35%), demonstrando um melhor equilíbrio para cenários onde o custo do falso alarme é alto.

3.3 Figuras

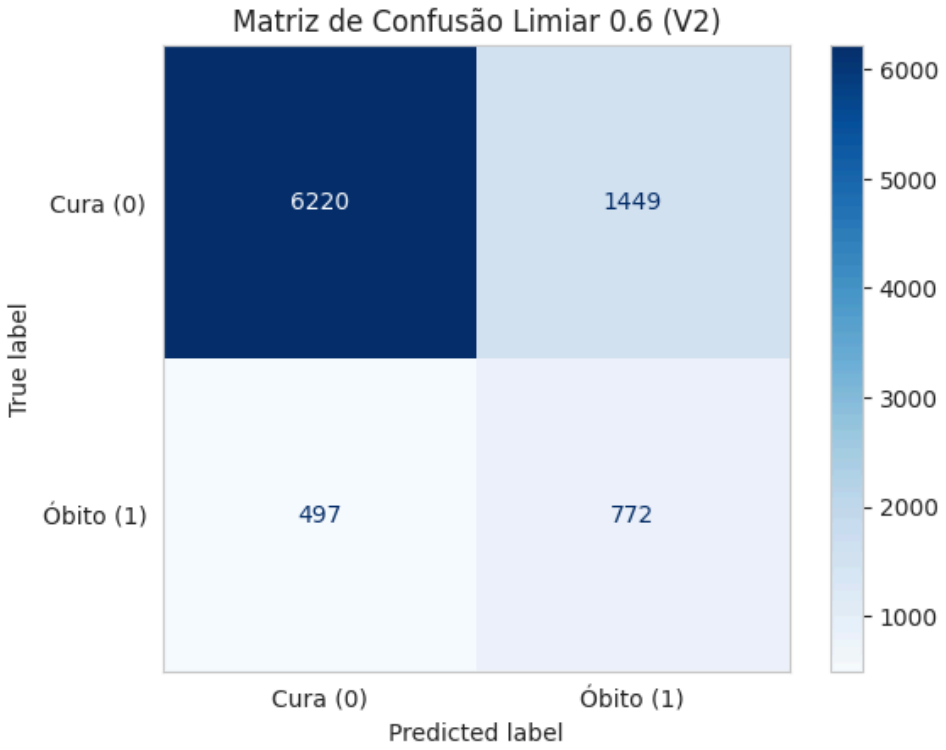
Abaixo estão as representações gráficas geradas durante a análise.



Histograma de Idade gerado na célula 11 do notebook *Figura 1. Distribuição de Idade: Cura vs Óbito, demonstrando a idade como fator de risco.*



Matriz de Confusão V1 gerada na célula 20 do notebook *Figura 2. Matriz de Confusão do Modelo V1 (Liniar 0,5), evidenciando alta sensibilidade mas baixa precisão.*



Matriz de Confusão V2 gerada na célula 22 do notebook] *Figura 3. Matriz de Confusão do Modelo V2 (Liniar 0,6), com redução de falsos positivos.*

4. Considerações Finais

O presente estudo atingiu seu objetivo principal ao demonstrar a viabilidade da utilização de algoritmos de Aprendizado de Máquina, especificamente a Regressão Logística, para a predição de óbito em pacientes diagnosticados com Síndrome Respiratória Aguda Grave (SRAG). A utilização de dados públicos do OpenDataSUS provou-se eficaz, permitindo a construção de um modelo com um poder de discriminação robusto (AUC-ROC de 0.79), mesmo diante dos desafios inerentes a dados reais, como o desbalanceamento de classes e valores ausentes.

Do ponto de vista clínico, a análise confirmou a idade como o fator de risco preponderante para o agravamento do quadro, corroborando a literatura médica e validando a coerência biológica do modelo. A comparação entre os limiares de decisão (0.50 na Versão 1 e 0.60 na Versão 2) revelou um *trade-off* estratégico fundamental para a aplicação hospitalar. O modelo original (V1) priorizou a sensibilidade (Recall de 72%), atuando como uma ferramenta de triagem "alarmista", ideal para cenários onde o custo de não identificar um paciente grave é inaceitável. Por outro lado, o modelo ajustado (V2) sacrificou parte dessa sensibilidade (queda de 11%) em troca de uma maior precisão e acurácia geral (78%), reduzindo significativamente a taxa de falsos positivos em 583 casos.

Conclui-se, portanto, que o modelo ajustado (limiar 0.60) apresenta-se como uma ferramenta clinicamente mais confiável para a gestão de recursos em ambientes de saúde saturados, pois é menos propenso a causar "fadiga de alertas" nas equipes médicas. Para trabalhos futuros, recomenda-se a exploração de técnicas avançadas de balanceamento de dados (como SMOTE) e a aplicação de métodos de *Ensemble* (como *Random Forest* ou *Gradient Boosting*), visando elevar a precisão das predições de óbito sem comprometer a capacidade de detecção dos casos críticos identificada neste estudo.

Referências

GRUS, Joel. Data science do zero: **noções fundamentais com Python.** 2. ed. Rio de Janeiro: Alta Books, 2016. Livro digital. (1 recurso online). ISBN 9788550816463. Disponível em: https://integrada.minhabiblioteca.com.br/books/9788550816463. Acesso em: 14 dez. 2025.