SIC 2022
X Seminário de Iniciação Científica do IFNMG
II Seminário da Pós-Graduação do IFNMG
8 a 10 de junho de 2022 Evento online
INSTITUTO FEDERAL
Norte de Minas Gerais
Campus Diamantina

DESENVOLVIMENTO DE UM MODELO DE SOBREVIDA EM PACIENTES COM SRAG POR COVID-19 NO BRASIL

BARROS, A.V.M.[1].; COSME, L.B.[2]; GUIMARÃES, V.H.D.[3];
[1]Discente do curso de Ciência da Computação do IFNMG – Campus Montes Claros; [2] Docente do IFNMG – campus Montes Claros; [3]Discente do Programa de Pós-graduação em Ciências da Saúde – PPGCS/Unimontes.

Palavras chaves: aprendizado de máquina, modelagem, regressão logística, tomada de decisão

Introdução

Descoberto em dezembro de 2019 na cidade de Wuhan (China), o vírus da COVID-19 se espalhou rapidamente pelo mundo, tornando-se um importante problema de saúde pública no mundo (HUANG, 2020). Até abril de 2022, já foram afetadas por volta de 500 milhões de pessoas pela doença em todo o mundo, ocasionando mais de 6 milhões de óbitos registrados.

Neste sentido, o monitoramento da evolução da doença no país, bem como a compreensão da dinâmica e evolução dos casos em pacientes por meio de modelos estatísticos, tem sido uma importante ferramenta para auxiliar na erradicação de doenças, além de auxiliar na tomada de decisões e condutas. Mediante a importância do exposto, o presente estudo teve por objetivo desenvolver um modelo preditivo de sobrevida e morte de pacientes com SRAG diagnosticados com COVID-19 no Brasil.

Material e métodos

Para esse estudo foram utilizados dados de domínio público, que possuem a licença Creative Commons Atribuição 4.0 Internacional (CC BY 4.0), e disponibilizados pelo OpenDataSUS. Os dados compreendem o período entre 22/07/2020 até 27/09/2021, totalizando 1.198.846 registros, sendo cada registro um caso da COVID-19 contendo informações como a semana de infecção, sintomas, comorbidades dos pacientes, bem como da anamnese. A partir do banco de dados foram filtrados casos de SRAG (Síndrome Respiratória Aguda Grave) por COVID-19, totalizando assim 706.093 casos. Ainda, foram selecionados casos como desfecho da SRAG por COVID-19, cura ou óbito, sumarizando 655.503 registros.

Devido a presença de valores nulos em várias colunas, optou-se por não imputar valores no processo de modelagem dos dados. Nesse caso, ao encontrar um valor nulo em uma coluna que fosse utilizada no processo de modelagem, toda a linha daquele registro com valor nulo era eliminada, tendo consequentemente um menor número de informações na base de dados final. Ao realizar essa filtragem, a base de dados teve uma redução de 95% do número de linhas, fato esse que, pode ter como consequência, uma perda da representatividade do modelo no cenário da doença, ficando composta portanto por 63.023 linhas. E quanto às colunas, por ser uma análise inicial da situação da doença, optou-se por considerar todas as colunas que poderiam ser relevantes para a predição, totalizando assim 35 colunas, sendo 34 delas utilizadas como variáveis preditoras para criar o modelo e 1 como a variável alvo, ou seja, aquele que será predita.

Ainda na triagem da amostra, com o intuito de proporcionar um melhor equilíbrio na predição das classes curados (n = 39.453) e óbitos (n = 23.570), utilizou-se a técnica de *undersampling*,

descrita por Herrera (2020). Que consiste basicamente em tentar igualar, ou aproximar o tamanho de um conjunto maior de dados para um conjunto menor, de modo que o conjunto de dados final não fique desbalanceado, ficando portanto com 47.140 registros para serem utilizados na modelagem.

Após a triagem dos dados, o trabalho se concentrou em três etapas: análise, modelagem e avaliação. Na etapa de análise, foram levantadas perguntas acerca dos dados e utilizou-se de estatística descritiva para melhor entender a base de dados. Para a etapa de modelagem, utilizou-se o algoritmo de aprendizado de máquina, Regressão Logística (em inglês, *Logistic Regressor*), incluído na biblioteca de código aberto *scikit-learn* da linguagem Python, versão 0.24.2. Para separação dos dados em treino e teste, utilizou-se a função *train_test_split*, também disponibilizada pela *scikit-learn*, sendo que o treino (n = 32.495) possuía aproximadamente 70% dos dados escolhidos de forma aleatória, e o teste (n = 14.645) com aproximadamente 30% dos dados.

E por último, na etapa de avaliação, examinou-se algumas métricas para avaliar e medir o desempenho do modelo encontrado. Para isso, foram utilizadas a acurácia e a especificidade do modelo, sendo que a acurácia se refere à soma dos acertos na classe dos curados e óbitos, dividido pelo total de predições realizadas, sendo portanto uma visão geral de como o modelo se performou. E a especificidade, apresenta a razão entre a predição dos óbitos que foi realizada corretamente e o número total de óbitos utilizado no modelo.

Resultados e discussão

O modelo construído teve uma acurácia de 74% e especificidade de 68%, ou seja, 68% dos óbitos foram preditos corretamente. Um importante ponto descoberto foi o elevado percentual de acerto para a classe dos curados, por volta de 80%, mesmo com o balanceamento de dados. Na Tabela 1 é possível visualizar a matriz de confusão do modelo, e com isso, verificar o quão ajustável o modelo se mostra para a classe dos curados. Olhando para os coeficientes do modelo de regressão logística, analisou-se que as informações com os coeficientes mais significativos foram, o uso do suporte ventilatório invasivo, que nesse caso, o paciente teria quase três vezes mais chances de ter o óbito como desfecho; a internação no hospital, ao ser internado o paciente seria 2 vezes mais promissor ao óbito; e por último a idade, verificou-se que a cada 20 anos mais velho, o paciente aumenta em 0.74 a chance de ter o óbito como desfecho.

Com a construção do modelo para a SRAG por COVID-19, foi possível identificar as características que mais impactaram para a classificação do desfecho dos infectados e consequentemente, que podem auxiliar na tomada de decisão sobre a doença. Um exemplo disso é a idade, variável que apresentou alta relevância para o modelo, e também é discutida nos trabalhos de Liu (2020) e Albitar (2020). Embora os trabalhos não especifiquem a análise da SRAG pela COVID-19, eles apresentam uma análise da COVID-19 de um modo geral, possibilitando assim correlacionar as informações do COVID-19 com a SRAG. Ainda, Liu em seu estudo apresenta o fator idade, como sendo um dos responsáveis pelo óbito entre os pacientes com COVID-19, mostrando a diferença significativa entre a idade dos curados e dos óbitos, o que também foi observado no presente estudo. Já o trabalho de Albitar apresenta uma análise de regressão multivariada para avaliar como alguns fatores contribuem para a mortalidade pela COVID-19. Albitar (2020) afirma que, “A idade parece ser o fator crucial para o resultado do COVID-19” e apresenta a hipertensão, doença cardiovascular crônica e doenças pulmonares também, como fatores importantes para a mortalidade pelo COVID-19.

Neste estudo, além da idade que foi identificada como um dos fatores mais importantes para a predição do desfecho da SRAG por COVID-19, visualizou-se também o uso do suporte ventilatório como fator decisivo para o resultado da SRAG por COVID-19, fato esse não apresentado pelos artigos discutidos anteriormente. De acordo com Marini (2020), o uso do suporte ventilatório poderia, em alguns casos, contribuir para o agravamento da situação dos pacientes com COVID-19, levando assim o infectado para óbito, fato esse que corrobora para validação do modelo construído.

Conclusão

O presente estudo ressalta a importância do cuidado com as pessoas em maior idade no enfrentamento da SRAG por COVID-19, visto que, à medida que a idade aumenta, o paciente tem mais chances de agravamento da doença. Além disso, é preciso investigar com mais detalhes o por quê a internação e o uso do suporte ventilatório, contribuíram de certa forma para o desfecho dos pacientes. E como perspectivas futuras, pode-se citar a busca por uma base de dados melhor, de modo a reduzir a fragilidade dos dados; o uso de técnicas de seleção de variáveis, permitindo que o modelo foque apenas nas variáveis mais promissoras; e por último, o uso de novos modelos, de forma a aprofundar e descobrir novas técnicas.

Agradecimentos

Este trabalho foi parcialmente apoiado pela Coordenação de Aperfeiçoamento do Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico, Tecnológico (CNPq) e Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Instituto Federal do Norte de Minas Gerais (IFNMG) Campus Montes Claros.

Referências

ALBITAR, O. et al. Risk factors for mortality among COVID-19 patients. **Diabetes research and clinical practice**. v.166, 2020.

Creative Commons. Atribuição 4.0 Internacional (CC BY 4.0). Disponível em: <https://creativecommons.org/licenses/by/4.0/deed.pt_BR#>. Acesso em: 12 Abr. 2022.

HERRERA, F. et al. **Learning from Imbalanced Data Sets**. Springer, 2018.

HUANG, C. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. **The Lancet**, v.395, n.10233, p.497-506, 2020.

LIU, W. et al. Analysis of factors associated with disease outcomes in hospitalized patients with 2019 novel coronavirus disease. **Chinese Medical Journal**, v.133, n.09, p.1032-1038, 2020.

MARINI J. J.; GATTINONI L. Management of COVID-19 Respiratory Distress. **JAMA**, 2020.

OpenDataSUS. SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. Disponível em: <https://opendatasus.saude.gov.br/dataset/srag-2020>. Acesso em: 12 Abr. 2022.

Scikit-Learn. Scikit-Learn Machine Learning in Python. Disponível em: <scikit-learn: machine learning in Python — scikit-learn 1.0.2 documentation>. Acesso em: 13 Abr. 2022.

ANEXO I

Tabela 1. Matriz de confusão

		Preditos	
		Curados	Óbitos
Reais	Curados	5898	1469
	Óbitos	2279	4999

Fonte: Arquivo Pessoal (2022).