ARTIGO

2965-6931 10.35699/2965-6931.2023.47705

REVISTA DA UNIVERSIDADE FEDERAL DE MINAS GERAIS

Estudo comparativo de modelos para previsão da mortalidade da SRAG por COVID-19 no Brasil

Luciana Balieiro Cosme
Instituto Federal do Norte de Minas Gerais (IFNMG)
https://orcid.org/0000-0002-3927-0329
luciana.balieiro@ifnmg.edu.br

André Vinícius Mendes Barros
Instituto Federal do Norte de Minas Gerais (IFNMG)
https://orcid.org/0000-0001-8356-372X
mb.andrevinicius@gmail.com

Victor Hugo Dantas Guimarães
Universidade Estadual de Montes Claros (UNIMONTES)
https://orcid.org/0000-0002-4424-8142
guimaraes.vhd@gmail.com

RESUMO

A síndrome respiratória aguda grave 2 (SARS-CoV-2) compreende uma das complicações desencadeadas pelo novo coronavírus. O presente trabalho tem como objetivo propor uma comparação entre dois modelos baseados em aprendizagem de máquina em diferentes contextos para prever a mortalidade nos casos da Síndrome Respiratória Aguda Grave (SRAG) pelo coronavírus 2019, COVID-19. Os dados utilizados estão disponíveis na plataforma DataSUS, e compreendem o período de janeiro de 2021 a dezembro de 2022. Consequentemente, realizou-se uma análise estatística descritiva, seleção de variáveis e, por fim, a elaboração de dois modelos, um antes do marco da segunda dose de vacinação para COVID-19 e outro depois. Em relação às métricas, o modelo 1 apresentou uma acurácia de 71,8%, enquanto o modelo 2 obteve uma acurácia de 80%, contribuindo portanto no processo de tomada de decisão para o enfrentamento da doença.

Palavras-chave: modelos preditivos; regressão logística; COVID-19.

Comparative study of mortality prediction models from (SARS) associated with COVID-19 in Brazil

ABSTRACT

Severe Acute Respiratory Syndrome 2 (SARS-CoV-2) comprises one of the complications triggered by the new coronavirus. This study aims to propose a comparison between two machine learning-based models in different contexts to predict mortality in cases of Severe Acute Respiratory Syndrome (SARS) caused by the 2019 coronavirus, COVID-19. The data used are available on the DataSUS platform and cover the period from January 2021 to December 2022. Consequently, a descriptive statistical analysis, variable selection, and the development of two models were carried out, one before the milestone of the second dose of COVID-19 vaccination and another after. Regarding the metrics, model 1 presented an accuracy of 71.8%, while model 2 achieved an accuracy of 80%, thus contributing to the decision-making process for tackling the disease.

Keywords: predictive models; logistic regression; COVID-19.

Submissão em: 15/08/2023 | **Aprovação em:** 12/11/2023

1. INTRODUÇÃO

Descoberto em dezembro de 2019 na cidade de Wuhan (China), o vírus de RNA envelopado de fita simples, pertencente ao gênero *Betacoronavirus* e família *Coronaviridae*, e nomeado como SARS-CoV-2 (Alexander E. Gorbalenya, 2020; de Groot R.J., 2020; Lu *al.*, 2020), se espalhou rapidamente pelo mundo e acabou se transformando em um grave problema de saúde pública global, segundo Huang *et al.* (2020). Por exemplo, comprovando as proporções do problema, dados mostraram que, em maio de 2022, mais de 528 milhões de pessoas em todo o mundo já haviam sido afetadas pela doença e mais de 6 milhões de óbitos haviam sido registrados. Em maio de 2023, porém, a Organização Mundial da Saúde (OMS) decretou o fim da Emergência de Saúde Pública de Importância Internacional referente ao SARS-CoV-2, mas, ainda assim, estudos sobre o tema são cada vez mais relevantes para que a doença possa ser melhor compreendida e seus efeitos negativos cada vez mais mitigados. Sob essa ótica, dados do DataSUS mostram que, nos casos de SARS-CoV-2 (COVID-19), a síndrome respiratória aguda é ainda imperativa para a mortalidade nos casos agravados (Li *et al.*, 2020). Além disso, estudos indicam uma taxa de mortalidade maior quando têm-se complicações como diabetes, doenças cardiovasculares, obesidade e doenças respiratórias prévias associadas (Cucinotta & Vanelli, 2020; Faria, 2021; Zhou *et al.*, 2021). Dentro dessa população, os idosos se destacam como um grupo de alto risco devido à subnutrição, baixa imunológica e diversas doenças crônicas associadas, que eles comumente podem apresentar. Desta forma, os dados disponíveis também apontam para uma importante questão de saúde pública no Brasil, tendo em vista a transição demográfica marcada pelo aumento de pessoas idosas, que corresponde a cerca de 14% (30,9 milhões) da população no país (Kuchemann, 2012; Carvalho & Rodriguez-Wong, 2008). Além disso, durante 2022, tornou-se evidente uma significativa redução no número de casos, especialmente após o alcance do marco de 70% da população brasileira vacinada com a segunda dose da vacina contra a COVID-19, registrado em 01/02/2022. Com o objetivo de analisar o impacto desse cenário pré e pós-vacinação, foram desenvolvidos dois modelos distintos.

O primeiro modelo abrangeu o período de 01/01/2021 a 01/02/2022, enquanto o segundo modelo englobou o período de 01/02/2022 a 31/12/2022. Esses modelos foram criados para investigar a influência da vacinação COVID-19 nos resultados e fornecer percepções sobre a eficácia das medidas de imunização em relação à SARS-CoV-2 e seus desdobramentos na saúde pública.

Neste sentido, o presente estudo tem como objetivo realizar uma comparação entre dois modelos baseados em aprendizado de máquina que utilizam dados sociodemográficos e fisiológicos para prever a mortalidade de pacientes com SRAG por COVID-19. O primeiro modelo é construído com dados antes do marco de 70% de vacinação da COVID-19, estimado em 01/02/2022, enquanto o segundo modelo é construído com dados após essa data.

2. REFERENCIAL TEÓRICO

O termo aprendizado de máquina é caracterizado como a área que estuda como usar computadores para simular atividades de aprendizado humano e estudar métodos de auto-aperfeiçoamento de computadores para obter novos conhecimentos e novas habilidades, identificar o conhecimento existente e melhorar continuamente o desempenho e sua realização, segundo Wang *et al.* (2009). Para esse trabalho, o aprendizado de máquina pode ser entendido como a área que reúne algoritmos que são capazes de simular atividades do aprendizado humano. Essa área possui três diferentes tipos de aprendizado: o aprendizado supervisionado, não supervisionado e por reforço.

O aprendizado supervisionado é um ramo do aprendizado de máquina, em que se tem como características o conhecimento dos vetores de entrada, valores que serão utilizados como entrada do modelo; e dos vetores de saída (Bishop, 2006), valores que serão utilizados como saída do modelo, o que se deseja prever. Esse tipo de aprendizado tem por intuito ajustar um modelo que relacione a resposta (vetor de saída) aos preditores (vetor de entrada). O objetivo é prever com precisão a resposta para observações futuras ou entender melhor a relação entre a resposta e os preditores (James

et al., 2013). Essa previsão ainda pode ser para: reconhecer uma determinada categoria de elementos em um conjunto de dados (algoritmos de classificação) ou prever um conjunto de valores nominais (algoritmos de regressão).

Formalmente, um problema de classificação possui: um conjunto de i entradas X = $\{X_1, X_2, ..., X_i\}$, um conjunto de j classes C = $\{c_1, c_2, ..., c_j\}$, sendo que, para cada conjunto X_i, uma ou mais classes ci serão seus rótulos; e por último, um conjunto de n vetores de treinamento T = $\{(x_1, \{c_1, ..., c_j\}), ...,(x_n, \{c_1, ..., c_n\})\}$. Com esses elementos, já é possível realizar o treinamento do modelo de classificação, por meio do uso do conjunto de vetores de treinamento que são inseridos para aprender ou estimar os parâmetros desconhecidos do modelo (Duda *et al.*, 2000). Ao identificar um conjunto de novas entradas, o modelo pode ser capaz de classificar a qual classe cada elemento do conjunto de entradas pertence. Esse estudo utiliza dois modelos de aprendizado de máquina, a regressão logística e a floresta aleatória.

2.1 REGRESSÃO LOGÍSTICA

Método paramétrico originado a partir da estatística e popular na área da saúde (Dreiseitl, 2002). Esse método é considerado paramétrico pois resume os dados com um conjunto de parâmetros de tamanho fixo, ou seja, caso tenha N variáveis $X_1, X_2, X_3, ..., X_n$, o modelo terá obrigatoriamente N + 1 parâmetros, conforme explicado por Norvig (2010).

A regressão logística tem como premissa ajustar os parâmetros do modelo de forma a minimizar a perda da soma dos resíduos dos quadrados no conjunto de dados (Norvig, 2010; David Die, 2022). A soma dos resíduos dos quadrados é dada pela Equação 1 e se refere à soma das diferenças entre o valor real (y_i) e o previsto ($\hat{y}_i$) ao quadrado. A vantagem de avaliar a regressão utilizando esse método é poder mensurar quão longe a resposta está do valor esperado.

$$SSE = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (1)$$

Para a previsão, a regressão logística faz o uso de um valor limiar entre 0 e 1 que, para valores que sejam maiores ou iguais a esse limiar, o modelo caracteriza como uma dada classe ou para valores abaixo desse limiar, como outra classe. Cria-se dessa forma, um classificador linear (Norvig, 2010). Ou seja, o que foi gerado pela função linear dada na Equação 2 é passado na função de limiar para verificar a qual classe aquele dado pertence. Na mesma equação o termo z é a própria função linear originada a partir do conjunto de dados. Assim, ao aplicá-la na função de limiar descrita, cria-se um classificador linear, sendo possível, portanto, realizar a classificação dos dados.

$$Logística(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Dentre os atributos que a função de regressão logística traz consigo, pode-se citar os valores dos coeficientes da regressão e a razão de chances (em inglês, *odds ratio*, OR). Os coeficientes da regressão são dados por números que representam o quanto uma dada variável será multiplicada pelo valor presente naquela variável para obtenção da previsão, possibilitando que um evento tenha probabilidade maior ou menor de ocorrer. Na Equação 3, β_0 e β_1 são os coeficientes das variáveis x_1 e x_2, respectivamente. Esses coeficientes são estimados, na maioria das vezes, pelo método de máxima verossimilhança, que possui a fórmula descrita na Equação 4. Segundo Subhash (1995), esse método busca estimar os parâmetros β_0 e β_1, de modo que, ao inserir essas estimativas no modelo para y, dado na Equação 3, produza um número próximo do objetivo real.

$$y = \frac{1}{1+e^{-(\beta_0 * x_1 + \beta_1 * x_2)}} \tag{3}$$

$$L(\beta_0, \beta_1) = \prod_{i:y_{i=1}}^{N} y_i * \prod_{i':y'_{i=0}}^{N} \left(1 - y'_i\right) \tag{4}$$

Conforme representado na Equação 5, a OR é uma medida de associação que traz consigo a razão entre duas probabilidades (Aguiar & Nunes, 2013), a probabilidade de ocorrência de um evento pela probabilidade de não ocorrência (cura). Por exemplo, suponha-se que a variável diabetes tenha um OR = 2, ou seja, ao dividir a probabilidade

de uma pessoa com diabetes morrer e a probabilidade de uma pessoa sem diabetes morrer, tem como resultado 2. Considera-se que, esse paciente que tem diabetes, possui 2 vezes mais chances de ir a óbito.

$$OR = \frac{\log log \frac{p_1}{1-p_1}}{\log log \frac{p_2}{1-p_2}} \quad (5)$$

Diferente da regressão logística que se baseia na estatística, a floresta aleatória (do inglês, *random forest*) é um método não-paramétrico que se baseia em múltiplas árvores de decisão. Ser considerado não-paramétrico se deve ao fato do modelo não ter suposições sobre a distribuição do espaço e a estrutura do classificador, ou seja, não se tem uma quantidade de parâmetros fixos como a regressão logística (Maimon & Rokach, 2010). Norvig (2010) descreve uma árvore de decisão como sendo uma representação para uma função que toma como entrada um vetor de valores de atributos e retorna uma decisão - de acordo com ele o retorno é de um único valor de saída. Em sua estrutura, a árvore de decisão é constituída por nós e ramificações, o primeiro nó recebe o nome de raiz e os últimos nós da parte baixa da árvore são chamados de nós folhas. Esse método funciona a partir das decisões tomadas em cada nó, particionando a entrada de modo que, ao final do particionamento, os nós folhas contenham elementos de uma única classe (Ratsch, 2004). Essa árvore tem como estrutura vários nós ligados de forma hierárquica: o nó raiz pode possuir filhos e cada um desses filhos outros filhos, assim sucessivamente. Para realizar a transição de um nó para o outro, deve ser verificada qual a decisão tomada naquele nó.

Como vantagens, podem ser citadas a sua fácil compreensão, possibilidade de ter dados de entrada nominais, numéricos e faltantes (Maimon & Rokach, 2010). E como desvantagens, a super adaptação (do inglês, *overfitting*), que é quando uma árvore se adapta muito a um determinado padrão durante o treinamento (Norvig, 2010).

O algoritmo da floresta aleatória segue o mesmo princípio de um algoritmo baseado em árvore de decisão, mas o modelo conta com diversas árvores construídas aleatoriamente de maneira a prover diversidade ao modelo (Lima *et al.*, 2021). Cada

árvore, ao receber um subconjunto aleatório do conjunto de dados, construirá a árvore de decisão e será capaz de classificar um dado conjunto de entrada. Como são múltiplas árvores de classificação, ao final da criação das árvores, a floresta adotará alguma estratégia para um problema em dados discretos, como por exemplo, escolher a árvore com mais votos e utilizá-la para classificar o conjunto de dados inserido (Sena, 2021; Breiman, 2001). Além do fato de ser um algoritmo de aprendizado supervisionado, a floresta aleatória pode ser utilizada como técnica para seleção de variáveis e, consequentemente, para redução da dimensionalidade do problema de modo a identificar quais variáveis foram mais importantes para a classificação (Sena, 2021). Na subseção seguinte serão apresentadas duas técnicas para seleção de variáveis, a seleção passo a passo para frente e o método da importância das variáveis (em inglês, *feature importances*) pela floresta aleatória, que busca classificar as variáveis por sua importância.

2.2 MÉTRICAS DE AVALIAÇÃO

A primeira técnica utilizada para auxiliar no processo de avaliação de um modelo é a validação cruzada. Faria (2021) define a validação cruzada como sendo uma ferramenta estatística utilizada para avaliar a capacidade de generalização de um modelo a partir de um determinado conjunto de dados não utilizado para obter as estimativas do desempenho do referido modelo. Sendo assim, com essa técnica é possível obter uma estimativa mais precisa do modelo e, caso se tenha mais de um, verificar aquele que se generalizou melhor aos dados (Norvig, 2010), apresentando-se como sugestão para a avaliação do desempenho de cada modelo. No método de validação cruzada com k repetições, a base de dados é dividida em k blocos iguais e, a cada rodada de treinamento, as informações referentes ao treino e teste mudam. Ao final dessas k repetições é realizada uma média do resultado de cada repetição do modelo. Um exemplo de resultado seria a quantidade de erros presentes nos dados de teste sendo possível reduzir a probabilidade de se ter pontuações muito boas ou muito ruins que não condizem com o modelo.

Para comparar os resultados de cada classe, têm-se as métricas de precisão, acurácia, sensibilidade e especificidade, que são bastante utilizadas em trabalhos relacionados com o tema. Já as métricas de desempenho se referem a quantidade de acertos a partir de um determinado ponto de vista. Para entender melhor sobre as métricas de desempenho de um modelo de aprendizado de máquina é preciso explicar os conceitos de verdadeiro positivo, falso positivo, verdadeiro negativo e falso negativo, a saber, segundo Provost (1998) e Franceschi (2019):

- Verdadeiro positivo (VP): número de óbitos que foram preditos como óbito;
- Falso positivo (FP): número de curados que foram preditos como óbito;
- Verdadeiros negativos (VN): número de curados que foram preditos como curados;
- Falso negativo (FN): número de óbitos que foram preditos como curado.

Quadro 1 - Matriz de confusão para o modelo de classificação.

		Reais	
		Óbitos	**Curados**
Preditos	**Óbitos**	VP	FP
	Curados	FN	VN

Fonte: elaborado pelo autor (2023)

A matriz de confusão de um modelo de classificação, representada no Quadro 1, é um instrumento importante pois apresenta diversas estatísticas originadas a partir de si, como por exemplo a acurácia, precisão e especificidade, possuindo portanto mais informações sobre o desempenho do modelo. Diante do contexto abordado, duas métricas podem ser priorizadas: a acurácia e a precisão, ambas utilizadas no aprendizado de máquina e descritas por Provost (1998).

> A acurácia, Equação 6, busca um equilíbrio entre os acertos dentro de duas classes (curados ou óbitos) apresentando o percentual de acertos sobre a variabilidade das predições realizadas. A precisão ou especificidade (em inglês, *precision*), Equação 7, informa quanto o modelo acertou da classe positiva ao analisar os valores preditos e a sensibilidade, dada na Equação 8, informa o quanto o modelo acertou os valores reais da classe positiva. Nesse último caso, busca-se analisar mais detalhadamente os valores reais da classe

positiva, classe que analisa os óbitos, pois é mais aceitável errar o desfecho de um paciente curado, que foi predito como óbito, do que o contrário.

$$acurácia = \frac{VP+VN}{VP+FP+FN+VN} \quad (6)$$

$$especificidade = \frac{VP}{VP+FP} \quad (7)$$

$$sensibilidade = \frac{VP}{VP+FN} \quad (8)$$

E a última métrica será a curva de característica de operação do receptor (AUCROC). Inicialmente, a curva ROC se baseia no cálculo da taxa de verdadeiros positivos (sensibilidade) e taxa de falsos positivos, conforme a Equação 8 e Equação 9 respectivamente. Essa curva se apresenta no gráfico em que o eixo X corresponde a taxa de falsos positivos e, o eixo Y corresponde a taxa de verdadeiros positivos, uma observação é que ambos os eixos possuem valor mínimo 0 e máximo 1. Com essa curva é possível analisar o desempenho de vários modelos com diferentes performances nas métricas apresentadas nessa seção e, assim, tomar a decisão de qual modelo escolher.

$$taxa\ de\ falsos\ positivos = 1 - especificidade \quad (9)$$

2.3 BALANCEAMENTO DE DADOS

Ao analisar o número de elementos em cada classe do trabalho, curados e óbitos, depara-se com um desbalanceamento das classes de maneira que o número de curados é muito maior que o número de óbitos, e consequentemente, uma classe (óbito) pode estar sub-representada no conjunto de dados (Herrera, 2018). No intuito de equilibrar as classes para que os modelos a reconheçam melhor, pode-se fazer o uso por exemplo de técnicas de balanceamento de classes baseadas em aprendizado sensível ao custo (Luo & Pan 2019).

Segundo Luo & Pan (2019), essa estratégia aborda o problema de desequilíbrio no nível de dados e no nível algorítmico, no qual um alto custo de classificação incorreta da classe minoritária é atribuído e o custo geral é minimizado. Sendo assim, o objetivo é ajustar os pesos de cada classe de modo que o modelo dê uma importância maior, as instâncias das classes minoritárias e o contrário para as classes majoritárias. Neste

trabalho, utiliza-se essa abordagem a nível de dados devido à simplicidade e já estar incorporado à biblioteca utilizada. No algoritmo de regressão logística disponível na biblioteca scikit-learn por exemplo, o balanceamento das classes pode ser feito automaticamente utilizando o parâmetro *class_weight* com o valor *'balanced'*. Por meio dessa configuração, o algoritmo ajusta os pesos automaticamente invertendo proporcionalmente a frequência de cada classe, conforme pode ser visto na Equação 10.

$$peso_{classe\ i} = \frac{Tamanho\ do\ conjunto\ de\ dados}{Quantidade\ de\ classes * Quantidade\ de\ amostras\ da\ classe\ i} \quad (10)$$

3. METODOLOGIA

Como tecnologias para o desenvolvimento, fez-se o uso da linguagem Python (https://www.python.org/) na versão 3.10.12, e das bibliotecas da própria linguagem como: Pandas (https://pandas.pydata.org/), para manipulação e análise dos dados; Numpy (https://numpy.org/pt/), para manipulação de operações aritméticas; scikit-learn (https://scikit-learn.org/stable/), para criação de modelos matemáticos e uso de métricas de avaliação dos modelos, e matplotlib (https://matplotlib.org/) e seaborn (https://seaborn.pydata.org/), para criação de gráficos. A base de dados para esse trabalho é de domínio público, sendo assim possuem a licença Creative Commons Atribuição 4.0 Internacional (CC BY 4.0) e são disponibilizados pelo openDataSUS (https://opendatasus.saude.gov.br/dataset/srag-2020), site que disponibiliza dados acerca da saúde no Brasil. Os dados compreendem o período de 01/01/2021 à 31/12/2022, totalizando 3.476.568 registros e uma prévia pode ser visualizada na Figura 1. Devido ao tema do trabalho, foram filtrados apenas os casos que tiveram como diagnóstico a SRAG por COVID-19 e selecionados os pacientes que tiveram desfecho, ou seja, foram curados ou morreram. Nesse trabalho, optou-se por não imputar valores, ou seja, caso a variável for preenchida pelo valor "Ignorado" ou estiver nula, a linha toda daquela informação foi retirada. Como consequência, obteve-se um menor número de informações na base de dados.

Figura 1 - Prévia do conjunto de dados utilizado.

CS_SEXO	NU_IDADE_N	CS_RACA	CS_ESCOL_N	FEBRE	TOSSE	GARGANTA	DISPNEIA	DESC_RESP	SATURACAO
F	56	1.0	9.0	1.0	1.0	1.0	2.0	1.0	2.0
F	67	1.0	9.0	1.0	1.0	2.0	1.0	1.0	1.0
M	68	1.0	9.0	1.0	1.0	2.0	2.0	2.0	2.0
F	42	2.0	9.0	1.0	1.0	2.0	1.0	1.0	1.0
M	45	1.0	NaN	1.0	1.0	2.0	1.0	1.0	2.0

Fonte: elaborada pelo próprio autor (2023)

4. RESULTADOS

Realizada a coleta do conjunto de dados utilizado neste trabalho, foram filtrados apenas os casos que tiveram como diagnóstico a SRAG por COVID-19, totalizando assim 3.017.888 casos. A base de dados contém informações como a semana de infecção, os sintomas, as comorbidades de pacientes, bem como da anamnese. Além disso, foram selecionados os pacientes que tiveram desfecho, ou seja, curados ou óbitos, resultando em 2.001.919 casos. Após a filtragem dos dados, percebeu-se a falta de informações principalmente em variáveis relacionadas com a presença de comorbidades como evidenciado no Quadro 2. O Quadro apresenta as colunas com as maiores quantidades de valores faltantes, os quais ultrapassam mais da metade do tamanho do conjunto de dados. Essa falta de dados prejudica a análise e as conclusões baseadas nessas informações, uma vez que essas colunas poderiam ser representativas e contribuir para o processo de modelagem. Nos trabalhos de Zhou *et al.* (2021), García *et al.* (2020) e Ovalle *et al.* (2021), por exemplo, são destacados que variáveis como diabetes e obesidade estiveram presentes nos modelos e se mostraram importantes para o agravamento da doença.

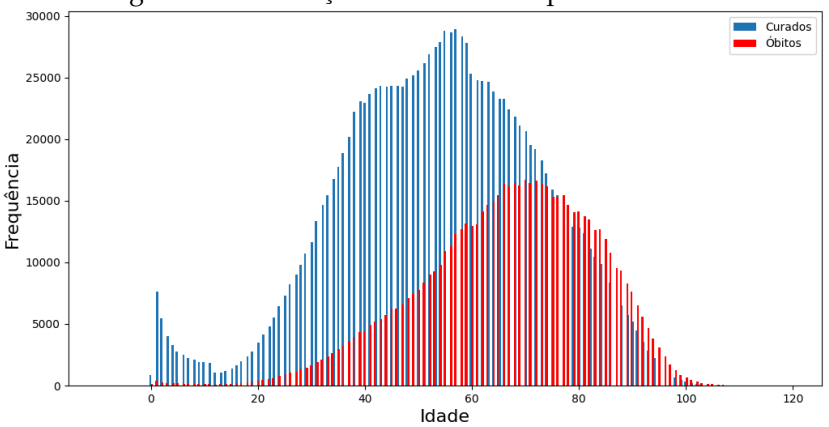
Quadro 2 - Quantidade e porcentagem de valores nulos das variáveis com os maiores percentuais.

Comorbidade [Nome da variável]	**Valores Nulos**	**% Valores Nulos**
Escolaridade [CS_ESCOL_N]	1.274.711	63,74%
Doença Hepática Crônica [HEPATICA]	1.272.820	63,64%
Doença Hematológica Crônica [HEMATOLOGI]	1.270.804	63,54%
Imunodeficiência [IMUNODEPRE]	1.263.254	63,16%
Asma [ASMA]	1.260.568	63,03%

Fonte: elaborado pelo próprio autor (2023)

Contudo, mesmo com a falta de informações, optou-se por não imputar valores, ou seja, caso a variável fosse preenchida pelo valor "Ignorado" ou estivesse nula, a linha toda daquela informação seria retirada, tendo como consequência um menor número de informações na base de dados. Logo, ao adotar esse critério, o conjunto de variáveis utilizado na etapa de modelagem não pode conter valores nulos. Analisando-se os dados dos infectados que possuíam desfecho da SRAG por COVID19, verificou-se que 66,625% (n=1.333.779) foram curados e 33,375% (n=668.140) faleceram. Observando-se o sexo dos infectados, foi verificado que o sexo masculino possui uma frequência maior, sendo 55,12% (n=1.103.476) contra 44,86% (n=898.211) feminino. Também foi registrado nessa variável o valor 'Ignorado', 232 vezes. Entre os casos por gênero, observou-se que 32,8% das mulheres foram registrados como óbito, enquanto esse número foi de 33,8% no sexo masculino dos desfechos analisados. Ao examinar a faixa etária dos infectados, observou-se que, aproximadamente, 70,9% dos infectados possuem entre 40 e 80 anos e a idade média dos óbitos foi de 66 anos, enquanto a dos curados foi de 53 anos. A visualização da distribuição por idade dos curados e óbitos pode ser observada na Figura 2. Visualmente, é possível notar que a concentração dos óbitos ocorre principalmente entre os indivíduos infectados com idade mais avançada, fato esse também mencionado por Albitar *et al.* (2020).

Figura 2. Distribuição dos Infectados por Idade.

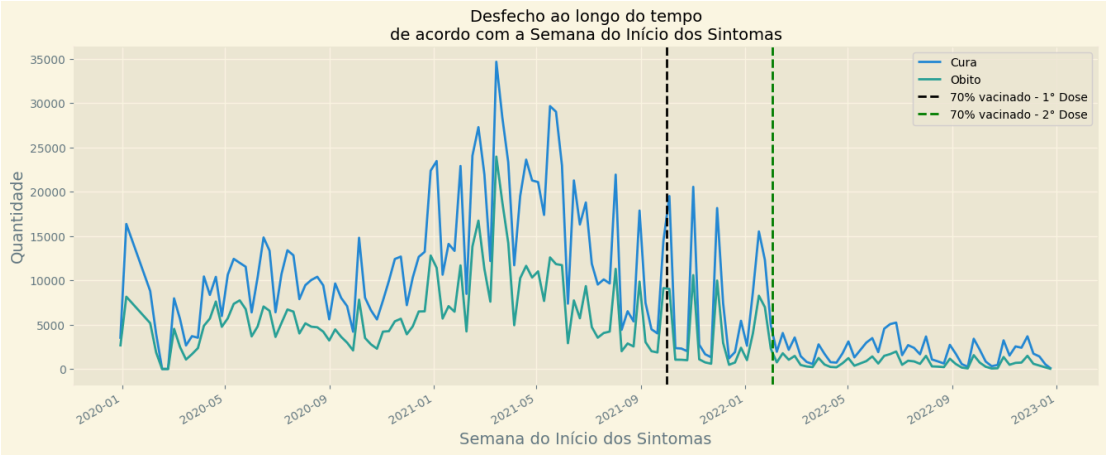


Fonte: elaborada pelo próprio autor (2023)

5. MODELAGEM

Na Figura 3 é apresentada a evolução do número de pessoas infectadas pela SRAG devido à COVID-19 ao longo das semanas desde o início da pandemia em 2020. A linha azul representa o número de casos de cura, enquanto a linha verde indica o número de óbitos registrados. É importante ressaltar que essa representação é um recorte limitado da realidade baseada nos dados disponíveis. Também é possível perceber que com o passar do ano de 2022, a quantidade de casos diminui significativamente, sobretudo, após marco, quando 70% da população brasileira estava vacinada com a segunda dose da vacina para COVID-19, estimado no dia 01/02/2022. Para observar o impacto do cenário antes e depois da vacinação, dois modelos foram elaborados, um compreendendo o período de 01/01/2021 até 01/02/2022 e outro de 01/02/2022 até 31/12/2022.

Figura 3 - Comparação entre a quantidade de curados e óbitos ao longo do tempo.



Fonte: elaborada pelo próprio autor (2023)

Com os períodos definidos, o próximo passo foi identificar quais variáveis seriam selecionadas. Para isso, utilizou-se o algoritmo de floresta aleatória para indicar a importância de cada variável para o modelo. O algoritmo recebe como entrada um conjunto de dados contendo informações sociodemográficas e fisiológicas, totalizando 31 variáveis que foram testadas. A saída dessa etapa teve como resultado uma lista ordenada decrescentemente da importância das variáveis, segundo o algoritmo. Posteriormente, essa lista de variáveis ordenadas foi adicionada em cada modelo de forma iterativa: uma variável é adicionada a cada iteração, testada com as demais e, caso

seja relevante para o modelo, é adicionada. Enquanto o algoritmo de floresta aleatória foi responsável por retornar a importância de cada variável, a regressão logística foi utilizada para de fato criar o modelo para prever a mortalidade dos infectados.

Para medir a relevância da variável em um modelo, foi verificado se a porcentagem de nulos dela não era maior que 15% da base de dados e se pelo menos duas destas três métricas: acurácia, recall ou AUCROC melhoram ao adicionar essa variável. Essas decisões foram feitas para os 2 modelos. O modelo de regressão logística que compreendia o período antes do marco de 70% de vacinados com a segunda dose da vacina para COVID-19, foi chamado de modelo 1. Esse modelo contou com 1.536.952 dados de treinamento e 61.270 de testes. Nesse modelo, utilizou-se as variáveis listadas onde seus coeficientes podem ser visualizados no Quadro 3:

- SUPORT_VEN_INV: Se utilizou o suporte ventilatório invasivo;
- NU_IDADE_N: Idade do infectado;
- SUPORT_VEN_NAO_INV: Se utilizou o suporte ventilatório não invasivo;
- SEM_PRI_TOTAL: Semana Epidemiológica do início dos sintomas desde o início da pandemia.

Quadro 3 - Razão de chance, coeficientes e intervalo de confiança para as variáveis do modelo 1.

Variável	Razão de Chances	Coeficiente	IC
SUPORT_VEN_INV	22,78	2,3203	2,308 – 2,332
SUPORT_VEN_NAO_INV	1,69	-0,1811	-0,018
NU_IDADE_N	1,04	0,0071	0,007 – 0,007
SEM_PRI_TOTAL	0,99	-0,0255	-0,001

Fonte: elaborado pelo próprio autor (2023)

Para o modelo 1 a acurácia foi de 71,8% e a área sob a curva ROC de 81,8%. No Quadro 4 é possível verificar que o modelo alcançou uma precisão de 55% para a classe de óbitos, o que significa que acertou corretamente 55% das predições feitas para essa classe. Além disso, obteve uma revocação de 76% para a classe de óbitos, o que indica que acertou 76% das instâncias que realmente eram óbitos. A revocação é uma métrica importante para o estudo, pois mede a capacidade do modelo em identificar

corretamente os casos positivos (nesse caso, os óbitos), fornecendo uma visão do quão bem o modelo está realizando essa tarefa específica.

Quadro 4 - Métricas do modelo do período antes do marco de 70% dos vacinados com a segunda dose da vacina para COVID-19.

Classe	Precisão	Revocação	Quantidade
Curados	0,85	0,7	41.100
Óbitos	0,55	0,76	20.170

Fonte: elaborado pelo próprio autor (2023.)

Para o modelo 2 também foi utilizado a regressão logística, esse modelo compreendia o período depois do marco de 70% dos vacinados com a segunda dose da vacina para COVID-19, dispondo de 102.710 dados na etapa de treinamento e 7.933 na etapa de teste. É possível destacar que a quantidade de dados presentes para treinar o modelo diminui significativamente devido ao corte no período dos dados. Quanto aos resultados, o modelo 2 teve as mesmas variáveis selecionadas do modelo anterior no método de seleção passo a passo para frente, mas com coeficientes do modelo diferentes, conforme apresentado no Quadro 5. Ao examinar as métricas no conjunto de teste, observou-se uma melhor acurácia (80%) e valores maiores nas demais métricas para a classe dos curados, enquanto o modelo 1 apresentou valores para as melhores métricas para a classe dos óbitos. Para completar, a métrica AUCROC foi de 80,9%, valor próximo do modelo 1. As métricas individuais de cada classe podem ser encontradas no Quadro 6.

Quadro 5 - Coeficientes, razão de chances e o intervalo de confiança para as variáveis do modelo 2.

Variável	Razão de Chances	Coeficiente	Intervalo de Confiança
SUPORT_VEN_INV	30,29	3,2695	3,217 - 3,322
SUPORT_VEN_NAO_INV	2,53	0,8441	0,804 - 0,884
NU_IDADE	1,03	0,0284	0,028 - 0,029
SEM_PRI_TOTAL	0,98	-0,0313	-0,032 - -0,031

Fonte: elaborado pelo próprio autor (2023)

Quadro 6 - Métricas do modelo do período depois do marco de 70% dos vacinados com a segunda dose da vacina para COVID-19.

Classe	Precisão	Revocação	Quantidade
Curados	0,9	0,74	3.611
Óbitos	0,45	0,71	1.068

Fonte: elaborado pelo próprio autor (2023)

O destaque de ambos modelos foi para as variáveis relacionadas ao uso do suporte ventilatório. O suporte ventilatório invasivo teve o valor de 22,78 para a razão de chances no modelo 1 e 30,29 para o modelo 2. Ou seja, caso o paciente precise utilizar esse suporte ventilatório, sua chance de óbito aumentaria consideravelmente mesmo após o período com 70% dos vacinados, evidenciando-se uma forte importância dessa variável neste trabalho. Fato esse destacado também no trabalho de MARINI & GATTINONI (2020), de acordo com o autor, o uso do suporte ventilatório pode piorar a condição de pacientes com COVID-19.

Vale ressaltar que essa piora pode ser por causa da própria situação do paciente, visto que o mesmo já se encontra em situação agravada por diversos fatores e necessita do procedimento mecânico para ajudar na respiração para sobreviver. Um dado interessante a se destacar é que 75% dos infectados após o marco de 70% dos vacinados com a segunda dose para COVID-19, que precisaram utilizar o suporte ventilatório invasivo morreram, reforçando a ideia de que não necessariamente seja o suporte ventilatório invasivo o principal causador das mortes, mas provavelmente, um estado já agravado por outros fatores anteriores e, que ao tentar utilizar o suporte ventilatório invasivo com esse quadro clínico, o paciente pode vir a desenvolver outras doenças como por exemplo a pneumonia e não resistir.

Outro destaque é para a variável de idade, com coeficientes de 1,04 e 1,03 para os modelos 1 e 2 respectivamente. Essa informação mostra que, a cada ano mais velho, a pessoa teria cerca de 4% mais chance de morrer antes do marco de 70% de vacinação para a segunda dose. Uma observação interessante é que, mesmo após essa data de referência da vacinação, o indivíduo infectado ainda teria um aumento de 3% na

probabilidade de óbito. Essa variável também foi importante e contribuiu para a mortalidade nos casos da COVID nos trabalhos de LIU *et al.* (2020) e ALBITAR *et al.* (2020). Embora esses trabalhos sejam referentes a COVID e não especifiquem a análise da SRAG causada pela COVID-19, é possível identificar a intersecção de ideias com o trabalho atual. Logo, destaca-se o cuidado com as pessoas mais idosas no enfrentamento da SRAG pela COVID-19 a fim de evitar o agravamento da doença. Além disso, no modelo 2 houve uma redução da razão de chances para cada ano na idade do paciente. Isto sugere uma importante mudança após o marco de vacinação, especialmente para as pessoas mais velhas. Em relação às métricas utilizadas para avaliar os modelos, verificou-se que a métrica AUCROC presente na literatura entre os trabalhos relacionados para a modelagem de dados da COVID-19 permeia em torno de 0,73 a 0,84 (Barda *et al.*, 2020). No presente trabalho a curva AUROC foi de 0,81 estando dentro da faixa de desempenho apresentada na literatura. Analisando-se a acurácia, percebe-se que alguns trabalhos relacionados tiveram métricas entre 72% e 82% (Hu *et al.*, 2021; Faria, 2021; Wollenstein-Betech *et al.*, 2020), contra 80% desse trabalho. Entretanto, vale ressaltar dois pontos no trabalho de Hu *et al.* (2021) que teve uma acurácia de 82%. Em primeiro lugar, o estudo utilizou variáveis laboratoriais que não estão presentes na base de dados utilizada. A inclusão dessas variáveis poderia melhorar os resultados, uma vez que fornecem informações adicionais sobre o paciente. Como segundo ponto, destaca-se o tamanho da base de dados utilizada para construir o modelo. Hu *et al.* (2021) utilizou 422 dados em seu modelo, enquanto para esse trabalho, o modelo 2 utilizou-se 102.710 dados para treinamento, que, consequentemente, proporciona maior robustez aos modelos aqui apresentados.

6. RESULTADOS

Conforme exposto, o presente estudo possibilitou a comparação entre os períodos pré-vacinação e pós-vacinação, bem como a identificação das características mais relevantes que impactaram na classificação da mortalidade em indivíduos afetados pela SRAG ocasionada pela COVID-19. Desse modo, mesmo diante de uma ampla

quantidade de dados ao longo do tempo e diversas variáveis, foi possível determinar o período e as variáveis mais importantes para construir os modelos que sejam mais próximos da situação atual, permitindo, consequentemente, a tomada de decisão mais rápida e efetiva. Uma contribuição significativa deste trabalho foi evidenciar o comportamento das variáveis selecionadas em relação ao marco de vacinação. Isto resultou em uma redução do impacto da idade e aumentou a importância do suporte ventilatório como fator determinante para o desfecho de óbito. Como limitação, o presente estudo aponta a ausência de dados clínicos e laboratoriais, refletindo na impossibilidade de fazer inferências acerca dos pacientes por meio de marcadores biológicos. Além disso, há um número considerável de dados faltantes nos formulários preenchidos, o que limita as análises. Apesar das limitações mencionadas, o presente estudo apresenta um tamanho amostral robusto, o que possibilita inferências confiáveis sobre os dados coletados e contribui para a tomada de decisões na conduta quanto aos pacientes. Como perspectivas para futuras pesquisas, pode-se citar o melhoramento da coleta de dados de forma a reduzir a fragilidade da base de dados utilizada e investigação mais aprofundada da relevância do suporte ventilatório invasivo no contexto da Síndrome Respiratória Aguda Grave causada pela COVID-19 nos casos de óbito.

7. CONSIDERAÇÕES FINAIS

Este trabalho foi parcialmente financiado pela Coordenadoria 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).

REFERÊNCIAS

AGUIAR, P.; NUNES, B. Odds Ratio: Reflexão sobre a Validade de uma Medida de Referência em Epidemiologia. **Acta medica portuguesa**, v. 26, n. 5, p. 505–510, 2013.

ALBERT, S.; LINVILLE, L. Benchmarking current and emerging approaches to infrasound signal classification. **Seismological research letters**, v. 91, n. 2A, p. 921–929, 2020.

ALBITAR, O. et al. Risk factors for mortality among COVID-19 patients. **Diabetes research and clinical practice**, v. 166, n. 108293, p. 108293, 2020.

BARDA, N. et al. **Performing risk stratification for COVID-19 when individual level data is not available – the experience of a large healthcare organization**. 2020. Disponível em: <http://dx.doi.org/10.1101/2020.04.23.20076976>.

BREIMAN, L. **Machine learning**, v. 45, n. 1, p. 5–32, 2001.

CARVALHO, J. A. M. DE; RODRÍGUEZ-WONG, L. L. A transição da estrutura etária da população brasileira na primeira metade do século XXI. **Cadernos de saude publica**, v. 24, n. 3, p. 597–605, 2008.

CHICCO, D.; JURMAN, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. **BMC medical informatics and decision making**, v. 20, n. 1, p. 16, 2020.

CORONAVIRIDAE STUDY GROUP OF THE INTERNATIONAL COMMITTEE ON TAXONOMY OF VIRUSES et al. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. **Nature microbiology**, v. 5, n. 4, p. 536–544, 2020.

CUCINOTTA, D.; VANELLI, M. WHO declares COVID-19 a pandemic. **Acta bio-medica : Atenei Parmensis**, v. 91, n. 1, p. 157–160, 2020.

DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. **Biometrics**, v. 44, n. 3, p. 837–845, 1988.

DIEZ, D.; ÇETINKAYA-RUNDEL, M.; BARR, C. **OpenIntro statistics**. [s.l: s.n.].

DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. **Journal of biomedical informatics**, v. 35, n. 5–6, p. 352–359, 2002.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 3. ed. [s.l.] Standards Information Network, 2022.

Elkan, C.P. (1997). **Boosting and Naive Bayesian learning**.

FARIA, A. R. Q. DE P. Análise de sobrevivência e fatores prognósticos associados à

mortalidade em pacientes com SRAG por Covid-19 hospitalizados em UTI na Paraíba. 2021.

FERNÁNDEZ, A. et al. **Learning from imbalanced data sets**. 1. ed. Basel, Switzerland: Springer International Publishing, 2018.

FERNÁNDEZ GARCÍA, L.; PUENTES GUTIÉRREZ, A. B.; GARCÍA BASCONES, M. Relationship between obesity, diabetes and ICU admission in COVID-19 patients. **Medicina Clínica (English Edition)**, v. 155, n. 7, p. 314–315, 2020.

FRANCESCHI, P. R. DE. Modelagens preditivas de Churn: o caso do Banco do Brasil. 2019.

GORBALENYA, A. E. et al. **Severe acute respiratory syndrome-related coronavirus: The species and its viruses – a statement of the Coronavirus Study Group**. 2020. Disponível em: <http://dx.doi.org/10.1101/2020.02.07.937862>.

GUDE-SAMPEDRO, F. et al. Development and validation of a prognostic model based on comorbidities to predict COVID-19 severity: a population-based study. **International journal of epidemiology**, v. 50, n. 1, p. 64–74, 2021.

HASTIE, T.; TIBSHIRANI, R.; TIBSHIRANI, R. J. **Extended comparisons of best subset selection, forward stepwise selection, and the lasso**. 2017. Disponível em: <http://arxiv.org/abs/1707.08692>.

HU, J.; FEI, Y.; LI, W.-Q. Predicting the mortality risk of acute respiratory distress syndrome: radial basis function artificial neural network model versus logistic regression model. **Journal of clinical monitoring and computing**, v. 36, n. 3, p. 839–848, 2022.

HUANG, C. et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. **Lancet**, v. 395, n. 10223, p. 497–506, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T. **An introduction to statistical learning: With applications in R**. New York, NY: Springer, 2013.

KÜCHEMANN, B. A. Envelhecimento populacional, cuidado e cidadania: velhos dilemas e novos desafios. **Sociedade e Estado**, v. 27, n. 1, p. 165–180, 2012.

LI, X. et al. Molecular immune pathogenesis and diagnosis of COVID-19. **Journal of pharmaceutical analysis**, v. 10, n. 2, p. 102–108, 2020.

LIMA, T. P. F. et al. Death risk and the importance of clinical features in elderly people with COVID-19 using the Random Forest Algorithm. **Revista Brasileira de Saúde Materno Infantil**, v. 21, n. suppl 2, p. 445–451, 2021.

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. 9, p. 1032–1038, 2020.

LU, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. **Lancet**, v. 395, n. 10224, p. 565–574, 2020.

LUO, H. et al. **Logistic regression and random forest for effective imbalanced classification**. 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). **Anais**...IEEE, 2019.

M., C. **Pattern Recognition and Machine Learning**. Nova Iorque, NY, USA: Springer, 2016.

MAIMON, O.; ROKACH, L. (EDS.). **Data mining and knowledge discovery handbook**. 2. ed. New York, NY: Springer, 2010.

MARINI, J. J.; GATTINONI, L. Management of COVID-19 respiratory distress. **JAMA: the journal of the American Medical Association**, v. 323, n. 22, p. 2329, 2020.

MARSONO, M. N.; EL-KHARASHI, M. W.; GEBALI, F. Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification. **Computer networks**, v. 53, n. 6, p. 835–848, 2009.

MONTAZERI, M. et al. Machine learning models in breast cancer survival prediction. **Technology and health care: official journal of the European Society for Engineering and Medicine**, v. 24, n. 1, p. 31–42, 2016.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to Linear Regression Analysis**. 3. ed. Nashville, TN: John Wiley & Sons, 2001.

MUSTAFA ABDULLAH, D.; MOHSIN ABDULAZEEZ, A. Machine learning applications based on SVM classification A review. **Qubahan Academic Journal**, v. 1, n. 2, p. 81–90, 2021.

OVALLE, D. L. P. et al. COVID obesity: A one-year narrative review. **Nutrients**, v. 13, n. 6, p. 2060, 2021.

PATIL, K. et al. **Deep learning based car damage classification**. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). **Anais**...IEEE, 2017.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Evaluating classifiers using ROC curves. **IEEE Latin America Transactions**, v. 6, n. 2, p. 215–222, 2008.

PROVOST, R. K. Glossary of terms. **Journal of Machine Learning**, v. 30, p. 271–274, 1998.

QUINLAN, J. R. Bagging, Boosting, and C4. 5. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 13, 1996.

RÄTSCH, G. A brief introduction into Machine Learning. 2004.

RIOS, L. F. B. **Modelos de predição de risco de morte para pacientes com carcinoma epidermoide de cabeça e pescoço**. [s.l.] Universidade de Sao Paulo, Agencia USP de Gestao da Informacao Academica (AGUIA), 2021.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: A modern approach**. [s.l.] Prentice Hall, 2010.

SENA, G. R. (2021). **Modelos Preditivos de Óbito para Pacientes de Óbito para Pacientes com COVID-19**.

SHARMA, S. **Applied Multivariate Techniques**. Nashville, TN: John Wiley & Sons, 1995.

TANBOĞA, I. H. et al. Development and validation of clinical prediction model to estimate the probability of death in hospitalized patients with COVID‐19: Insights from a nationwide database. **Journal of medical virology**, v. 93, n. 5, p. 3015–3022, 2021.

WANG, H.; MA, C.; ZHOU, L. **A brief review of machine learning and its application**. 2009 International Conference on Information Engineering and Computer Science. **Anais**...IEEE, 2009.

WOLLENSTEIN-BETECH, S. et al. Physiological and socioeconomic characteristics predict COVID-19 mortality and resource utilization in Brazil. **PloS one**, v. 15, n. 10, p. e0240346, 2020.

ZHOU, Y. et al. Obesity and diabetes as high-risk factors for severe coronavirus disease 2019 (Covid-19). **Diabetes/metabolism research and reviews**, v. 37, n. 2, p. e3377, 2021.