Inteligência artificial e previsão de óbito por Covid-19 no Brasil: uma análise comparativa entre os algoritmos *Logistic Regression, Decision Tree* e *Random Forest*

Artificial intelligence and forecasting of death by COVID-19 in Brazil: A comparative analysis of the algorithms Logistic Regression, Decision Tree, and Random Forest

Risomario Silva[1], Darcy Ramos da Silva Neto[2]

DOI: 10.1590/0103-11042022E809

RESUMO Este trabalho fez uso da inteligência artificial para contribuir com evidências empíricas que auxiliem na previsão de morte por Covid-19, possibilitando a melhoria de protocolos de saúde utilizados em sistemas de saúde no Brasil e dotando a sociedade com mais ferramentas de combate a essa doença. Utilizaram-se dados de janeiro a setembro de 2021 para o Brasil com o objetivo de prever morte por Covid-19, tomando por base o quadro clínico de pacientes que utilizaram o Sistema Único de Saúde no período estudado. Três algoritmos de classificação foram experimentados: *Logistic Regression* (LR), *Decision Tree* (DT) e *Random Forest* (RF). Os modelos LR, DT e RF tiveram uma acurácia média de, respectivamente, 76%, 76% e 77% na previsão de morte. Além disso, foi possível inferir que, quando o paciente chega a um ponto que necessita do uso de suporte ventilatório e de Unidade de Terapia Intensiva, somado à idade, sua chance de ir a óbito por Covid-19 é maior.

PALAVRAS-CHAVE Covid-19. Sars-CoV-2. Modelos logísticos. Inteligência artificial. Aprendizado de máquina.

ABSTRACT *This work makes use of artificial intelligence to contribute with empirical evidence that help predict death by COVID-19, enabling the improvement of health protocols used in health systems in Brazil and providing society with more tools to combat COVID-19. Data from January to September 2021 for Brazil are used in order to predict death by COVID-19 based on the clinical status of patients who used the Unified Health System in the studied period, in which three classification algorithms were tried: Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF). The LR, DT, and RF models had a mean accuracy of 76%, 76%, and 77% in predicting death, respectively. In addition, it was possible to infer that when patients reach a point that require the use of ventilatory support and ICU, added to age, their chance of dying of COVID-19 is greater.*

KEYWORDS *COVID-19. SARS-CoV-2. Logistic models. Artificial intelligence. Machine learning.*

[1] Universidade Federal de Pernambuco (UFPE) – Recife (PE), Brasil.
risomario.c.e@hotmail.com

[2] Universidade de São Paulo (USP) – Ribeirão Preto (SP), Brasil.

Introdução

Apesar de habituados a viver de uma maneira mais restrita – com *lockdowns*, restrições físicas, toques de recolher, entre outros mecanismos –, a pandemia causada pelo Sars-CoV-2 ainda é um desafio a ser enfrentado pela sociedade. Não obstante o avanço da vacinação em diferentes partes do mundo, constantemente, noticiários e jornais relatam novas ondas de infecções a todo momento.

Dados da World Health Organization[1] estimam que mais de 6,57 milhões de pessoas perderam suas vidas para a Covid-19 no globo desde o começo da pandemia, onde 625 milhões de pessoas já foram infectadas. Com cerca de 34,7 milhões de pessoas infectadas e 687 mil mortes por Covid-19, até a escrita final desta pesquisa, o Brasil aparentava estar enfrentando uma fase de estabilidade: dados das secretarias estaduais de saúde mostram que o País apresentou uma média móvel de mortes diárias de 72 e uma média móvel de 5,6 mil novas infecções para o décimo segundo dia de outubro de 2022.

Um dos maiores problemas enfrentados no começo da pandemia foi a falta de informação acerca do novo vírus, pois a dificuldade em estabelecer protocolos de saúde bem definidos atrasava o trabalho de profissionais e de equipes de saúde na classificação de prioridades no atendimento de pacientes infectados. Com o avanço contínuo de infecções no mundo todo, os bancos de dados passaram a ser fomentados, permitindo uma maior precisão de informações e protocolos de saúde mais adequados.

O surto da Covid-19 afetou seriamente a forma como a sociedade vive, gerando crises econômicas mundialmente, alterando o padrão de comportamento e a saúde humana. Com a dinâmica da epidemia global se tornando cada vez mais séria, a previsão e a análise de casos e mortes por Covid-19 tornaram-se uma tarefa importante para pesquisadores.

Embora muito conhecimento acerca da Covid-19 no Brasil tenha sido adquirido, ainda existe bastante dúvida quanto à letalidade de grupos populacionais distintos. Assim, estudos se esforçaram em classificar as principais razões que potencializam o risco de morte por Covid-19 no Brasil[2-4]. No entanto, parte desses trabalhos tem problemas em comum: amostra pouco representativa, geralmente associada a um estado ou região específica, perdendo em aleatoriedade; ou análise estatística fraca, carecendo de modelos estatísticos mais robustos.

Nesse sentindo, esta pesquisa tem o objetivo de desenvolver modelos de classificação que possam auxiliar na previsão de morte por Covid-19, a partir quadro clínico do paciente, utilizando-se de inteligência artificial por meio de três algoritmos de classificação: *Logistic Regression*, *Decision Tree* e *Random Forest*. Isso será feito com base em informações do quadro clínico de cerca de 134.639 mil pacientes diagnosticados com essa doença que passaram pelo Sistema Único de Saúde (SUS) entre janeiro e setembro de 2021, em várias regiões do Brasil. Algoritmos inteligentes orientados por aprendizado de máquina podem contribuir e capacitar respostas eficientes à pandemia da Covid-19 ao melhorar modelos de prognóstico rotineiramente usados em clínicas no mundo. Isso pode ajudar a prever os resultados de saúde da Covid-19 em diversos ambientes geográficos e de sistemas de saúde.

O aprendizado de máquina em campos como os da bioinformática, da saúde planetária e da tomada de decisão clínica em geral está em um momento crítico; sua capacidade de examinar em tempo real conjuntos de dados altamente diversos pode ajudar a construir resiliência em sistemas de saúde planetários em resposta às pandemias presentes e futuras, sendo útil também na alocação de ajuda para gestores de sistemas de saúde planetários e equipes multiprofissionais[5,6].

Material e métodos

Como relatado na introdução deste trabalho, será feito uso de três algoritmos específicos: *Logistic Regression*, *Decision Tree* e *Random Forest*. A seguir, os três algoritmos serão explicitados em detalhes.

Logistic Regression (Regressão Logística)

A regressão logística[7-9] é um modelo linear para classificação. Também é conhecida na literatura como regressão *logit*, classificação de entropia máxima ou classificador log-linear. A regressão logística binária ou univariada representa os casos de regressão logística em que a variável dependente é binária ou dicotômica, isto é, assume apenas dois valores. Nesse caso, a variável dependente Y segue uma distribuição de *Bernoulli*, tendo uma probabilidade desconhecida *p*. Na regressão logística, é feita a estimação da probabilidade desconhecida, dada uma combinação linear das variáveis independentes.

Como Y segue uma distribuição de *Bernoulli*, é preciso um *link* para ligar essa distribuição presente em Y às variáveis independentes: essa ligação é chamada de função *logit*. A razão de probabilidades é chamada de *odds*, e seu logaritmo natural, o *logit*, é a função de ligação:

$$\ln(odds) \rightarrow ln\left(\frac{p}{1-p}\right) \quad (1)$$

Seja a função linear das variáveis independentes dada por:

$$g(x) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} \quad (2)$$

Em que **β** é o vetor de estimadores e **X** é a matriz de variáveis independentes. Das equações (1) e (2), tem-se:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} \quad (3)$$

Como o objetivo é estimar *p*, precisa-se resolver (3), de modo a obter (6):

$$\frac{p}{1-p} = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} \quad (4)$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}+\cdots+\beta_k X_{k,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}+\cdots+\beta_k X_{k,i}}} \quad (5)$$

$$\hat{p} = \frac{e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}+\cdots+\beta_k X_{k,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}+\beta_2 X_{2,i}+\cdots+\beta_k X_{k,i}}} \quad (6)$$

A equação (6) é chamada de equação de regressão estimada e representa o objetivo do modelo de regressão logística dado que $\hat{p}$ é a probabilidade estimada de um determinado evento acontecer, óbito ou recuperação por exemplo.

Decision Tree (Árvore de Decisão)

Decision Tree[7-9] é um método de aprendizado supervisionado, não paramétrico, usado para classificação e regressão. O objetivo é construir um modelo que preveja o valor de uma variável de destino (óbito ou recuperação), aprendendo regras de decisão simples inferidas a partir de um conjunto de dados. Dados vetores de treinamento $X_i \in R^n$, i=1, ... L e um vetor de classe $y \in R^L$, uma árvore de decisão particiona recursivamente o espaço de *Features* de modo que as amostras com as mesmas classes ou valores de destino semelhantes sejam agrupadas.

Sejam os dados no nó *m* representados por Q_m com N_m amostras. Para cada divisão $\theta = (j, t_m)$ consistindo de uma *Feature j* e um limite t_m, particione o conjunto de dados em $Q_m^{left}(\theta)$ e $Q_m^{right}(\theta)$ conjuntos:

$$Q_m^{left}(\theta) = \left\{(x,y) \,\middle|\, x_j \leq t_m\right\} \quad (7)$$

$$Q_m^{right}(\theta) = Q_m \setminus Q_m^{left}(\theta) \quad (8)$$

A qualidade de uma divisão candidata do nó *m* é então calculada usando uma função de impureza ou função de perda *H()*, escolha que depende da tarefa a ser resolvida (classificação ou regressão):

$$G(Q_m, \theta) = \frac{N_m^{left}}{N_m} H(Q_m^{left}(\theta)) + \frac{N_m^{right}}{N_m} H(Q_m^{right}(\theta)) \quad (9)$$

Selecione os parâmetros que minimizam a impureza:

$$\theta^* = argmin_\theta G(Q_m, \theta) \quad (10)$$

Repita para os subconjuntos em $Q_m^{left}(\theta^*)$ e $Q_m^{right}(\theta^*)$ até que a profundidade máxima permitida seja alcançada $N_m < min_{a\text{-}mostras}$ ou $N_m = 1$.

Existem algumas vantagens e desvantagens no uso de árvores de decisão: é simples de entender e de interpretar, requer pouca preparação de dados, é capaz de lidar com dados numéricos e categóricos, consegue lidar com problemas de múltiplas saídas, usa um modelo de caixa branca, isto é, se uma determinada situação é observável em um modelo, a explicação para a condição é facilmente explicada pela lógica booleana; por fim, é possível validar um modelo por meio de testes estatísticos.

As desvantagens no uso de *Decision Tree* são: *overfitting*, as árvores de decisão podem ser instáveis porque pequenas variações nos dados podem resultar na geração de uma árvore completamente diferente, as previsões das árvores de decisão não são suaves nem contínuas, mas aproximações constantes por partes; por fim, o modelo pode criar árvores tendenciosas se algumas classes dominam.

Random Forest (Floresta Aleatória)

Random Forest[7-9] é um metaestimador que ajusta vários classificadores de *Decision Tree* em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e o controle de *overfitting*. Em *Random Forest*, cada árvore no conjunto é construída a partir de uma amostra retirada com substituição (amostra de *bootstrap*) do conjunto de treinamento.

Em geral, o modelo *Decision Tree* geralmente exibe alta variação e tende a se ajustar demais. A aleatoriedade injetada no *Random Forest* produz árvores de decisão com erros de previsão um tanto dissociados. Tirando uma média dessas previsões, alguns erros podem ser cancelados. O *Random Forest* alcança uma variação reduzida combinando diversas árvores, às vezes ao custo de um ligeiro aumento no viés. Na prática, a redução da variância é frequentemente significativa, resultando em um modelo geral melhor

Dado um conjunto de treinamento $X = x_1, x_2, ..., x_n$ com respostas $Y = y_1, y_2, ..., y_n$, é feito o ensacamento (*bagging*) repetidamente (K vezes) selecionando uma amostra aleatória com substituição do conjunto de treinamento e se ajustam árvores para essas amostras.

Para K = 1,...k, uma amostra com substituição n exemplos de treinamento X, Y, são determinados: X_k,Y_k. Em seguida, uma *Decision Tree* é treinada, f_k em X_k,Y_k. Após o treinamento, as previsões para novas amostras x' podem ser realizadas a partir da média das previsões de todas as árvores individuais sobre x':

$$\hat{f} = \frac{1}{K}\sum_{k=1}^{K} \hat{f}_k(x') \quad (11)$$

Este procedimento leva a um melhor desempenho do modelo, pois reduz a variância dele. Isto implica que enquanto as previsões de uma única *Decision Tree* são altamente sensíveis ao ruído no seu conjunto de treinamento, a média de muitas delas não é, desde que as árvores não sejam correlacionadas.

Dados

Este trabalho se utilizou do Banco de Dados de Síndrome Respiratória Aguda Grave (SRAG), do Ministério da Saúde[10], no qual foram filtrados apenas pacientes que testaram positivo para Covid-19 entre janeiro e setembro de 2021. O Ministério da Saúde desenvolve a vigilância da SRAG no Brasil desde a pandemia de Influenza A(H1N1)pdm09 em 2009. Em 2020, a vigilância da Covid-19 foi incorporada na rede de vigilância da Influenza e outros vírus respiratórios. Após realizado o procedimento de mineração e estruturação da base de dados, um total de 134.639 observações compõe a amostra deste estudo.

Resultados

As estatísticas descritivas apontam um total de 134.639 pessoas, sendo 51,84% de homens e com idade média de 60 anos. A *tabela 1* também apresenta o quantitativo de comorbidades na

distribuição dessa amostra, os sintomas apresentados e as taxas de utilização de Unidade de Terapia Intensiva (UTI), suporte ventilatório e percentual de vacinados nesse período.

Tabela 1. Estatísticas clínicas

Total = 134.639		
Variáveis	**n**	**%**
Sexo		
Masculino	69.809	51,84
Feminino	64.830	48,16
Idade média	**60**	
Comorbidades		
Cardiopatia	70.110	52,07
Doença Hematológica	1.211	0,89
Doença Hepática	1.552	1,15
Doença Neurológica	6.837	5,07
Doença Renal	6.378	4,73
Asma	5.585	4,15
Diabetes	47.064	34,95
Pneumopatia	5.924	4,40
Imunodepressão	4.337	3,22
Obesidade	26.429	19,62
Sintomas apresentados		
Fadiga	51.029	37,90
Perda de olfato	15.845	11,76
Perda de Paladar	16.375	12,16
Febre	61.902	45,97
Inflamação na Garganta	25.975	19,29
Dispneia	106.834	79,34
Desconforto respiratório	89.828	66,71
Saturação menor que 95%	27.545	20,45
Evolução		
Número de óbitos	54.279	40,31
Número de recuperados	80.360	59,68
UTI	**52.827**	**39,23**
Suporte Ventilatório	**99.671**	**74,02**
Vacina	**50.066**	**37,18**

Fonte: elaboração própria com base nos dados do Banco de Dados de Síndrome Respiratória Aguda Grave – SRAG/Ministério da Saúde[10].

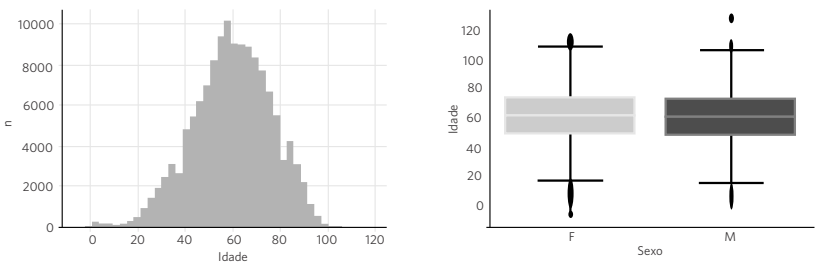
A *tabela 1* explicita as variáveis trabalhadas neste artigo e algumas características da amostra utilizada. Percebe-se que, entre o grupo de pacientes aqui estudado, cerca de 52% deles são cardiopatas, 35% são diabéticos e 19% são obesos; as demais comorbidades se apresentam em menor grau. Os pacientes apresentam mais frequentemente sintomas como fadiga (37,90%), febre (45,97%), dispneia (79,34%) e desconforto respiratório (66,71%).

Para além disso, cerca de 40% dos pacientes fizeram uso de UTI, e 74,02% deles se utilizaram de suporte ventilatório (invasivo ou não). No tocante à evolução dos casos, um total de 59,68% dos pacientes da amostra se recuperou, enquanto 40,31% foram a óbito.

É importante frisar que a taxa de vacinação (pelo menos uma dose) entre os sobreviventes e os que foram a óbito é bastante similar, situando-se em torno de 37%, ou seja, um percentual relativamente baixo para que se permita traçar relações de causa e efeito entre número de vacinados e taxa de mortalidade no período de referência desta pesquisa. É válido salientar que o Brasil iniciou a vacinação contra a Covid-19 em meados de janeiro de 2021; e a passos muito lentos, em junho do mesmo ano, apenas 12,41% dos brasileiros haviam tomado as duas doses da vacina.

Nota-se ainda que cerca de 52% dos indivíduos da amostra são homens e que 48% são mulheres, a média de idade dos pacientes da amostra é de 60 anos. O *gráfico 1* a seguir apresenta a distribuição de idade da amostra e a distribuição da idade por sexo.

Gráfico 1. Distribuição da idade da amostra



Fonte: elaboração própria com base nos dados do Banco de Dados de Síndrome Respiratória Aguda Grave – SRAG/Ministério da Saúde[10].

Aqui é importante traçar algumas discussões acerca da distribuição da idade dos indivíduos da amostra: a média de idade dos sobreviventes é de 56,66, enquanto a idade média dos indivíduos que foram a óbito é de 65 anos, uma diferença superior a 8 anos, refletindo o fato de que pacientes infectados com idade mais avançada têm um risco de morte mais elevado quando comparados a pacientes mais novos.

A *tabela 2* a seguir traz uma síntese das principais métricas e resultados dos modelos aqui estimados. Nesse sentido, uma matriz de confusão foi construída (matriz na qual se visualizam os acertos e os erros do modelo); e, a partir de seus resultados, foram obtidas algumas métricas.

Tabela 2. Avaliação de desemprenho dos classificadores

Evolução	Precisão	F1-score	Recall
Logístic Regression			
Cura	0.75	0.81	0.89
Óbito	0.78	0.66	0.57
Acurácia			0.76
AUC ROC			0.73

Tabela 2. Avaliação de desemprenho dos classificadores

Evolução	Precisão	F1-score	Recall
Decision Tree			
Cura	0.77	0.81	0.86
Óbito	0.76	0.68	0.62
Acurácia			0.76
AUC ROC			0.74
Random Forest			
Cura	0.78	0.82	0.87
Óbito	0.76	0.69	0.63
Acurácia			0.77
AUC ROC			0.75

Fonte: elaboração própria com base nos dados do Banco de Dados de Síndrome Respiratória Aguda Grave – SRAG/Ministério da Saúde[10].

A acurácia representa um percentual total de acertos do modelo. Essa métrica nem sempre é uma medida muito boa para trabalhar com modelos de classificação, pois ela pode induzir a achar que um modelo que prediz corretamente uma classe A, mas erra muito ao predizer a classe B seja um modelo muito bom. Por isso, outras métricas foram consideradas. A precisão é a capacidade do modelo de não prever uma instância negativa como positiva (não cometer erro do tipo 1), ou seja, para todas as instâncias classificadas como positivas, qual é o percentual de acerto.

Já a métrica *recall* mostra a capacidade do modelo de encontrar todas as instâncias positivas, isto é, para todas as instâncias que são, de fato, positivas, qual é o percentual de acerto. A métrica F1, por sua vez, conjuga as duas anteriores como uma média harmônica entre ambas. Uma excelente alternativa é fazer a *Receiver Operating Characteristic* (ROC) e calcular a *Area Under the Curve* (AUC). A curva ROC mensura a capacidade de predição do modelo proposto por meio das predições da sensibilidade e da especificidade. Essa técnica serve para visualizar, organizar e classificar o modelo com base na performance preditiva[11]; em termos práticos, quanto mais ao noroeste do *gráfico 2* a curva estiver, melhor:

Gráfico 2. AUC-ROC

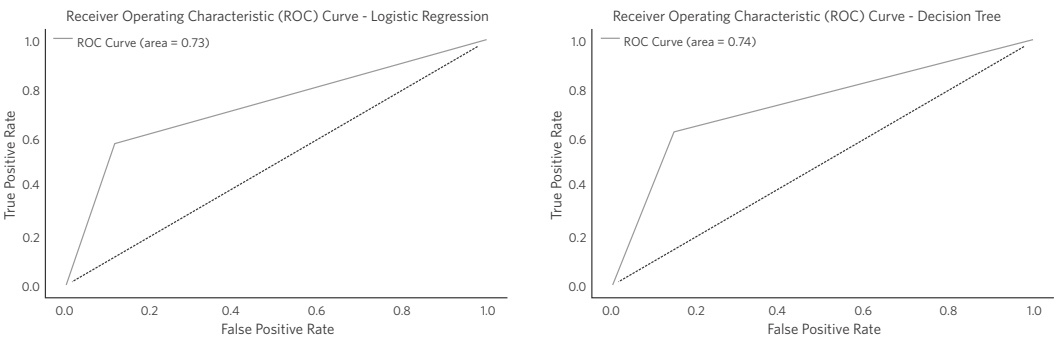
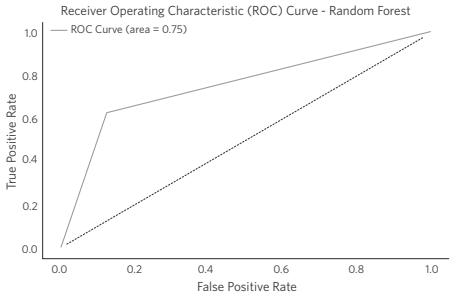


Gráfico 2. AUC-ROC



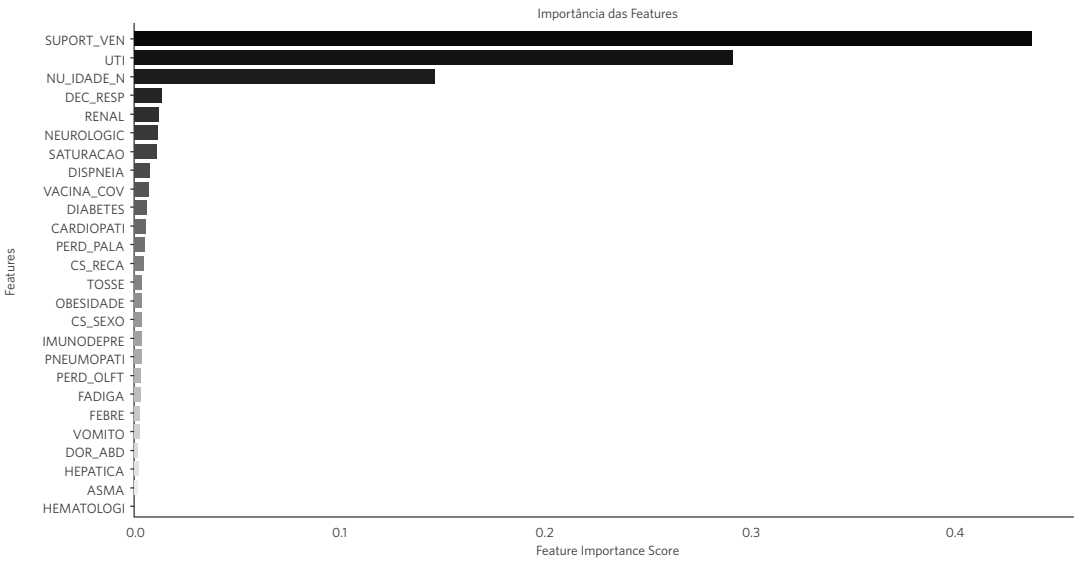
Fonte: elaboração própria com base nos dados do Banco de Dados de Síndrome Respiratória Aguda Grave – SRAG/Ministério da Saúde[10].

Nota-se que o classificador *Logistic Regression* prevê morte por Covid-19 com uma acurácia de 0.76 e uma precisão de 0.78; além disso, apresenta uma área sob a curva ROC de 0.73 conforme o *gráfico 2*. O classificador *Decision Tree* é capaz de prever morte por Covid-19 com uma acurácia de 0.76 e uma precisão de 0.76; ademais, apresenta uma área sob a curva ROC de 0.74. Para o classificador *Decision tree*, foi realizada uma série de simulações (poda da árvore) de modo a evitar *overfitting*. Por fim, o classificador *Random Forest* consegue prever morte por Covid-19 com uma acurácia de 0.76 e uma precisão de 0.77; outrossim, apresenta uma área sob a curva ROC de 0.75, conforme o *gráfico 2*. Foi realizado o *tunning* do modelo no intuito de adotar os hiperparâmetros mais adequados de modo a evitar *overfitting*; no mais, o modelo foi validado com *Cross Validation*.

Observa-se que as métricas apresentadas pelos três modelos analisados foram muito similares, mas isso não é uma informação redundante; muito pelo contrário, reflete a robustez dos resultados que foram encontrados. Como os resultados do *Random Forest* foram ligeiramente melhores que os demais classificadores, a discussão em cima da importância das *Features* será feita a partir de seus resultados. Como pode ser visto no *gráfico 3*, a importância das *Features* calculada por esse algoritmo mostrou que uso de suporte ventilatório (0.46), uso de UTI (0.27), Idade (0.14) e Demais comorbidades (0.13), sendo as doenças renais, neurológicas, diabetes, cardiopatia e obesidade como as mais relevantes, são as principais variáveis que explicam a evolução do paciente a óbito por Covid-19.

É importante atentar ao seguinte fato: o uso de UTI e de suporte ventilatório é variável relevante para separar sobreviventes dos não sobreviventes, por isso seu elevado nível de importância atribuído pelo algoritmo; mas isso não implica dizer que essas *Features* sejam as causadoras dessas mortes, muito pelo contrário, pacientes com estados clínicos críticos são mais propensos a fazer uso desses instrumentos e, portanto, têm uma probabilidade maior de ir a óbito.

Gráfico 3. Importância das Features



Fonte: elaboração própria com base nos dados do Banco de Dados de Síndrome Respiratória Aguda Grave – SRAG/Ministério da Saúde[10].

Discussão

A partir dos resultados encontrados, é possível inferir que pacientes que apresentam um quadro clínico severo, associado à presença de comodidades, estão mais propensos a fazer uso de mecanismos como suporte ventilatório e UTI. Portanto, eles têm um maior risco de morte por Covid-19. Apesar de usar metodologia diferente, o resultado aqui encontrado comunga com a literatura[12], que aponta que cerca de 80% dos pacientes intubados no Brasil entre 16 de fevereiro e 15 de agosto de 2020 foram a óbito.

Embora o Brasil tenha um elevado número de hospitais e leitos de UTI, quando comparado a outros países da Europa, há uma elevada heterogeneidade na distribuição regional destes[13-16]. Além disso, a falta de um protocolo nacional que unifique as técnicas utilizadas e o uso de pessoal sem treinamento e experiência adequados foram apontados como causas importantes das mortes por Covid-19 no Brasil[12].

Somado a isso, a superlotação de hospitais leva a uma demora na intubação de pacientes graves, piorando o quadro clínico. Pelo menos até março de 2021, não havia por parte do Ministério da Saúde um protocolo de atuação para intubação de infectados. Tempo e recursos foram gastos discutindo tratamentos precoces sem qualquer evidência científica, e não se investiu em disseminar informação sobre tratamentos eficazes para pacientes graves, como uso de esteroides, técnicas de identificação de insuficiência respiratória, uso da posição prona, entre outros[12].

A idade se mostrou como um importante determinante do quadro final do paciente com Covid-19 (conforme *gráfico 3*). Como discutido, o grupo de pacientes que foram a óbito tinha uma idade média de quase nove anos acima do grupo dos sobreviventes, o que corrobora outros estudos que também mostram a idade como um importante atributo para explicar morte por Covid-19. A literatura já chegou a um acordo comum de que a idade é uma importante característica para explicar o desfecho de um paciente contaminado com Covid-19. Trabalhos indicam que, no Brasil, na China, no México e na Europa, mortes por Covid-19 estavam concentradas em uma faixa etária acima dos 60 anos. Isso pode estar associado

ao fato de esse grupo ter mais comorbidades associadas quando comparado à população mais jovem. Além disso, como já abordado, o Brasil foi retardatário em elaborar um plano de ação nacional de combate à Covid-19, com estados e regiões definindo suas próprias políticas públicas de atuação contra o vírus[2,17–21].

Considerações finais

Este estudo estimou modelos de classificação que auxiliam na previsão de morte por Covid-19 no Brasil, usando dados de janeiro e setembro de 2021. Os modelos estimados apresentaram métricas moderadas. O modelo que se apresentou como mais adequado foi o classificador *Random Forest*, com uma acurácia de 77 % e AUC-ROC de 75%. Nesse sentindo, o modelo foi capaz de prever morte por Covid-19 e revelar as comorbidades mais importantes nessa previsão: idade, doenças renais, doenças neurológicas, diabetes, cardiopatia e obesidade, que levam, por sua vez, a uma maior necessidade de uso de instrumentos como suporte ventilatório e UTI, elevando a probabilidade de morte dos indivíduos.

Pode-se enfatizar o uso de algoritmos como uma ferramenta adicional para a sociedade no combate à pandemia da Covid-19, permitindo que pesquisadores de saúde planetária vão além do emprego de relações lineares entre um número restrito de observações e contribuam para a redução de erros de diagnóstico e o uso de ferramentas de diagnóstico ineficientes.

A pandemia da Covid-19 mostrou que a saúde digital é possível, viável e não é uma realidade distante da sociedade. O papel do aprendizado de máquina na medicina e na saúde continuará a crescer, principalmente após o choque de tecnologia causado pela pandemia, em que cenários futuristas foram adiantados e já se encontram materializados e disponíveis para uso da população.

As principais dificuldades desta pesquisa estão relacionadas com o Banco de Dados de SRAG, o qual está repleto de informações faltantes, contribuindo, assim, para redução da amostra após o tratamento dos dados. No mais, para trabalhos futuros, sugere-se a separação dos grupos por faixa etária e regiões no intuito de identificar se fatores demográficos como infraestrutura de saúde e variáveis sociais, por exemplo, contribuem ou não para o aumento de mortes por Covid-19.

Assim sendo, este trabalho atinge o objetivo proposto de prever morte por Covid-19 para o Brasil, baseando-se no quadro clínico dos pacientes que passaram pelo SUS. Do melhor conhecimento dos autores, este é um dos poucos trabalhos que se utilizam de algoritmos de classificação para realizar previsão de óbito por Covid-19 no Brasil com base no quadro clínico do paciente. Desse modo, esta pesquisa contribui com evidências empíricas que fundamentam a adoção de protocolos de saúde adequados e a construção de políticas públicas eficientes para treinamento e atuação de profissionais no processo de recuperação de pacientes infectados pelo Sars-Cov-2, no intuito de reduzir o número de consequências fatais.

Colaboradores

Silva R (0000-0003-3908-2923)* e Silva Neto DR (0000-0003-4864-8167)* contribuíram igualmente para a elaboração do manuscrito. ■

*Orcid (Open Researcher and Contributor ID).

Referências

1. World Health Organization. Coronavirus (COVID-19) Dashboard: 2022. Geneva: World Health Organization; 2022. [acesso em 2022 mar 9]. Disponível em: https://covid19.who.int.

2. Sena GR, Lima TPF, Vidal SA, et al. Clinical Characteristics and Mortality Profile of COVID-19 Patients Aged less than 20 years Old in Pernambuco–Brazil. Am J Trop Med Hyg. 2021 [acesso em 2022 mar 28]; 104(4):1507-1512. Disponível em: https://www.ajtmh.org/view/journals/tpmd/104/4/article-p1507.xml.

3. Lima TPF, Sena GR, Neves CS, et al. Previsão de óbito e importância de características clínicas em idosos com COVID-19 utilizando o Algoritmo Random Forest. Rev. Bras. Saude Mater. Infant. 2021 [acesso em 2022 mar 20]; 21(supl2):445-451. Disponível em: https://www.scielo.br/j/rbsmi/a/C65mNw8xgR6Td6LKkCzxzkN/abstract/?lang=pt#:~:text=Resultados%3A,satura%C3%A7%C3%A3o%20de%20oxig%C3%AAnio%20%E2%89%A495%25.

4. Galindo RJSC, Andrade LB, Sena GR, et al. Mulheres com câncer e COVID-19: uma análise da letalidade e aspectos clínicos em Pernambuco. Rev. Bras. Saude Mater. Infant. 2021 [acesso em 2022 mar 15]; 21(supl1):157-165. Disponível em: https://www.scielo.br/j/rbsmi/a/Tp3Xnrwmt5DC49ZvNZBQjpG/?lang=pt.

5. Blumenstock J. Machine learning can help get COVID-19 aid to those who need it most. Nature. 2020 maio 14. [acesso em 2022 mar 15]. Disponível em: https://www.nature.com/articles/d41586-020-01393-7#:~:text=14%20May%202020-,Machine%20learning%20can%20help%20get%20COVID%2D19%20aid%20to%20those,share%20lessons%20and%20minimize%20risks.

6. Arga KY. COVID-19 and the Futures of Machine Learning. Omics. J.Iintegr. biol. 2020 [acesso em 2022 mar 15]; 24(9):512-514. Disponível em: https://www.liebertpub.com/doi/10.1089/omi.2020.0093.

7. Hastie T, Tibshirani R, Fridman, J. The elements of statistical learning. New York: Springer; 2009.

8. Breiman L, Friedman J, Olshen RA, et al. Classification and regression trees. Abingdon: Routledge; 2017.

9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. The J. machine Learning research. 2011 [acesso em 2022 mar 17]; 12(85):2825-2830. Disponível em: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

10. Brasil. Ministério da Saúde. SRAG 2020 - Banco de Dados de Síndrome Respiratória Aguda Grave - incluindo dados da COVID-19. 2021. [acesso em 2022 mar 15]. Disponível em: https://opendatasus.saude.gov.br/dataset/srag-2020-banco-de-dados-de-sindrome-respiratoria-aguda-grave-incluindo-dados--da-covid-19.

11. Fawcett T. An introduction to ROC analysis. Pattern recognition letters. [acesso em 2022 maio 15]; 27(8):861-874. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X.

12. Ranzani OT, Bastos LSL, Gelli JGM, et al. Characterisation of the first 250000 hospital admissions for COVID-19 in Brazil: a retrospective analysis of nationwide data. The Lancet Resp. Med. 2020 [acesso em 2022 maio 15]; 9(4):407-418 Disponível em: https://www.thelancet.com/journals/lanres/article/PIIS2213-2600(20)30560-9/fulltext.

13. Austin S, Murthy S, Wunsch H, et al. Access to urban acute care services in high- vs. middle-income countries: an analysis of seven cities. Intens. Care Med. 2013 [acesso em 2022 maio 2]; 40(3):342-352. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3938845/.

14. Salluh JIF, Lisboa T. Critical care in Brazil. ICU Manag. Pract. 2016 [acesso em 2022 maio 2]; 16(3):188-191. Disponível em: https://healthmanagement.org/c/icu/issuearticle/critical-care-in-brazil-1.

15. Machado FR, Cavalcanti AB, Bozza FA, et al. The epidemiology of sepsis in Brazilian intensive care units (the Sepsis PREvalence Assessment Database, SPREAD): an observational study. Lancet Infect Dis. 2017 [acesso em 2022 maio 2]; 17(11):1180-89. Disponível em: https://pubmed.ncbi.nlm.nih.gov/28826588/.

16. Azevedo LCP, Park M, Salluh JIF, et al. Clinical outcomes of patients requiring ventilatory support in Brazilian intensive care units: a multicenter, prospective, cohort study. Critical Car. 2013 [acesso em 2022 maio 2]; 17(2):1-13. Disponível em: https://pubmed.ncbi.nlm.nih.gov/23557378/.

17. World Health Organization. COVID-19: WHO european region operational update epi weeks 31–32 (27 July–9 August). Geneva: World Health Organization; 2020. [acesso em 2022 maio 2]. Disponível em: https://www.euro.who.int/__data/assets/pdf_file/0008/460196/COVID-19-operational-update-weeks-31-32-eng.pdf.

18. Wu D, Wu T, Liu Q, et al. The SARS-CoV-2 outbreak: what we know. Inter. J. Infect. Diseases. 2020 [acesso em 2022 maio 2]; (94):44-48. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32171952/.

19. Bello-Chavolla OY, González-Díaz A, Antonio-Villa NE, et al. Unequal impact of structural health determinants and comorbidity on COVID-19 severity and lethality in older Mexican adults: Considerations beyond chronological aging. The J. Geront. Series A. 2021 [acesso em 2022 maio 25]; 76(3):52-59. Disponível em: https://pubmed.ncbi.nlm.nih.gov/32598450/.

20. Brasil. Ministério da Saúde, Secretaria de Vigilância em Saúde. Doença pelo coronavírus COVID-19: semana epidemiológica (16 a 22/08). Boletim Epidemiol. 2020; (34).

21. Costa JA, Silveira JA, Santos SCM, et al. Implicações cardiovasculares em pacientes infectados com Covid-19 e a importância do isolamento social para reduzir a disseminação da doença. Arq. Bras. Cardiol. 2020 [acesso em 2022 abr 2]; 114(5):834-838. Disponível em: https://www.scielo.br/j/abc/a/YLLdXBRX7zjhtFVgmhKsjQF/?lang=pt#:~:text=Com%20o%20objetivo%20de%20mostrar,e%20espanhol%2C%20dispon%C3%ADveis%20na%20plataforma.

Recebido em 30/05/2022
Aprovado em 24/10/2022
Conflito de interesses: inexistente
Suporte financeiro: não houve