Gustavo Cardozo Rodrigues

Modelo preditivo para prognóstico de pacientes com COVID-19

Alegrete

2021

Gustavo Cardozo Rodrigues

Modelo preditivo para prognóstico de pacientes com COVID-19

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal do Pampa como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal do Pampa

Orientador: Prof. Dr. Diego Kreutz

Coorientador: Prof. Dr. Mirkos Ortiz Martins

Alegrete

2021

Ficha catalográfica elaborada automaticamente com os dados fornecidos
pelo(a) autor(a) através do Módulo de Biblioteca do
Sistema GURI (Gestão Unificada de Recursos Institucionais) .

R696m Rodrigues, Gustavo Cardozo
Modelo preditivo para prognóstico de pacientes com COVID-19 / Gustavo Cardozo Rodrigues.
47 p.

Trabalho de Conclusão de Curso(Graduação)-- Universidade Federal do Pampa, CIÊNCIA DA COMPUTAÇÃO, 2021.
"Orientação: Diego Luis Kreutz".

1. Aprendizado de máquina. 2. Priorização de pacientes. 3. Classificação de risco. 4. Prognóstico COVID-19. I. Título.

Gustavo Cardozo Rodrigues

Modelo preditivo para prognóstico de pacientes com COVID-19

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Bacharel em Ciência da Computação.

Monografia defendida e aprovada em: 30/06/2021.

Banca examinadora:

_____

Prof. Dr. Diego Kreutz

Orientador

(UNIPAMPA)

_____

Prof. Dr. Mirkos Ortiz Martins

Coorientador

(UFN)

Prof. Dr. Marcelo Resende Thielo
(UNIPAMPA)

Prof. Dr. Adriano Velasque Werhli
(FURG)

Prof. Alex Camargo
(APUS Digital)

Assinado eletronicamente por **ADRIANO VELASQUE WERHLI**, **Usuário Externo**, em 30/06/2021, às 20:17, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Mirkos Ortiz Martins**, **Usuário Externo**, em 30/06/2021, às 20:18, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **DIEGO LUIS KREUTZ**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 30/06/2021, às 20:28, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Alex Dias Camargo**, **Usuário Externo**, em 30/06/2021, às 20:30, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **MARCELO RESENDE THIELO**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 01/07/2021, às 18:40, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0558836** e o código CRC **4DC73F91**.

Dedico este trabalho a Deus, por ser essencial em minha vida, meu socorro bem presente na hora da tribulação. Porque dele, e por ele, e para ele são todas as coisas; glória, pois, a ele eternamente. Amém!

AGRADECIMENTOS

A Deus, pela minha vida, e por ter me sustentado até aqui.

Aos meus pais, Claudia e Paulo, aos meus irmãos Cezar e Matheus, a minha vó Nara, e a minha namorada Gabriela pelo amor e incentivo.

Aos amigos Edinei Silva e Fernanda Souza, pelo apoio em dias trabalhosos.

Aos professores, pelas correções e ensinamentos que me permitiram apresentar um melhor desempenho no meu processo de formação profissional. Em especial aos amigos e orientadores, Dr. Diego Kreutz e Dr. Mirkos Martins, que acreditaram em mim, e me ajudaram a extrair o meu potencial. Muito obrigado pela paciência e pelos "puxões de orelha" que me deram! O meu profundo e eterno agradecimento.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

"O que adquire entendimento ama a sua alma;
o que cultiva a inteligência achará o bem."
(BÍBLIA, Provérbios, 19, 8).

RESUMO

Com a criticidade da falta de leitos de unidade de terapia intensiva durante a pandemia de COVID-19, políticas para determinar quem tem acesso aos leitos foram implantadas em estados do Brasil. Objetivando mitigar a subjetividade dessas políticas, propomos o desenvolvimento e validação interna de um preditor para classificação de risco de óbito de pacientes com COVID-19 no estado do Rio Grande do Sul. Utilizamos o guia TRIPOD para explanar o desenvolvimento do modelo de aprendizagem de máquina baseado em floresta aleatória. O conjunto de dados possui 604.389 registros e engloba os pacientes atendidos no Rio Grande do Sul que foram reportados através do Painel Coronavírus RS, no perído de 01 de janeiro à 08 de junho de 2021. A variável de desfecho (*outcome*) é a evolução, que informa se o paciente foi recuperado ou veio a falecer por COVID-19. No total, quatorze características foram elencadas como preditoras, sendo elas demográficas (sexo e faixa etária) e clínicas (sintomas e comorbidades). O conjunto de dados de derivação possui 408.959 registros, sendo 3,18% (n=13.005) óbitos, e o conjunto de validação possui 175.269 registros, sendo 5.574 óbitos. No conjunto de testes, o modelo classificou a chance de óbito com uma pontuação AUC-ROC de 0,981 (intervalo de confiança de 95% 0,981 a 0,982) e AUC-ROC de 0,970 no conjunto de validação. O classificador permite estratificar com boa precisão o risco de óbito de pacientes com COVID-19 e pode contribuir na dimuinição da subjetividade na tomada de decisão no ambiente hospitalar. Entretanto, salientamos que esta é apenas uma ferramenta adicional, e todos os aspectos éticos e legais devem ser considerados na tomada de decisão médica.

Palavras-chave: Aprendizado de Máquina. Priorização de Pacientes. Classificação de Risco. Prognóstico COVID-19.

ABSTRACT

With the criticality of the lack of beds in the intensive care unit during the COVID-19 pandemic, policies to determine who has access to beds were implemented in Brazilian states. Aiming to mitigate the subjectivity of these policies, we propose the development and internal validation of a predictor for death risk classification of patients with COVID-19 in Rio Grande do Sul. We use the TRIPOD guideline to explain the development of the random forest-based machine learning model. The dataset has 604,389 records and includes patients treated in Rio Grande do Sul who were reported through the Painel Coronavírus RS, in the period from January 1 to June 8, 2021. The outcome variable is evolution, which informs whether the patient was recovered or died from COVID-19. In total, fourteen characteristics were listed as predictors, which were demographic (gender and age group) and clinical (symptoms and comorbidities). The derivation dataset has 408,959 records, being 3.18% (n=13,005) deaths, and the validation set has 175,269 records, being 5,574 deaths. In the test set, the model rated the chance of death with an AUC-ROC score of 0.981 (95% confidence interval 0.981 to 0.982) and AUC-ROC of 0.970 in the validation set. The classifier makes it possible to stratify with good precision the risk of death of patients with COVID-19 and can contribute to the reduction of subjectivity in decision-making in the hospital environment. However, we point out that this is just an additional tool, and all ethical and legal aspects must be considered in decision-making.

Key-words: Machine Learning. Patient Priorization. Risk Stratification. COVID-19 Prognosis.

SUMÁRIO

1 INTRODUÇÃO

Durante a pandemia de *coronavirus disease 2019* (COVID-19) os sistemas de saúde do mundo inteiro enfrentam problemas quanto a disponibilidade e alocação de recursos como respiradores e leitos de UTI (Unidade de Terapia Intensiva) (LATIF et al., 2020) (RANNEY; GRIFFETH; JHA, 2020). Políticas e decisões críticas estão sendo tomadas quanto a priorização de pacientes com a doença e, em países como o Brasil, existem regras que definem quem tem direito a leitos de UTI. Em alguns lugares, devido a urgência e falta de dados mais precisos, a política tem sido priorizar as pessoas mais jovens para a ocupação dos leitos de UTI[1]. Em estados como o Rio Grande do Sul, não há leitos para todos os pacientes em estado grave[2], consequentemente, é preciso colocar em prática políticas de priorização da alocação dos leitos com base no estado clínico do paciente.

Estudos mostram que modelos de aprendizagem de máquina (*machine learning*) conseguem predizer a chance de óbito de um paciente positivo para COVID-19 com até 0,99 de pontuação AUC-ROC (*Area Under the Curve - Receiver Operating Charachteristic Curve*) (WYNANTS et al., 2020). Por se tratar de um contexto crítico (pandemia e risco de vida), os modelos preditivos devem ser robustos, seguros e transparentes, sem comprometer a ética. Um dos principais desafios é desenvolver e garantir a credibilidade destes modelos.

Muitos estudos e propostas de soluções para classificação de risco refletem dificuldades nas suas aplicações ou até mesmo em como podem auxiliar no contexto médico (WYNANTS et al., 2020). Diversos modelos contém alto risco de enviesamento (*i.e.*, modelo não generalista) devido ao desenvolvimento baseado em amostras pequenas da população (*e.g.*, dados de apenas um hospital), métricas de avaliação inadequadas (*e.g.*, acurácia para dados desbalanceados), ou pobreza de detalhes sobre o desenvolvimento e objetivo do modelo (*e.g.*, omissão dos métodos utilizados) (WYNANTS et al., 2020). Devido a alta variabilidade climática, cultural e racial entre diferentes países e populações (ASSAF et al., 2020; CHENG et al., 2020; ZHAO et al., 2020; CASIRAGHI et al., 2020) é necessário o desenvolvimento de uma solução que utilize dados de populações pouco exploradas, como é o caso do Brasil, além de utilizar métricas para avaliação do desempenho de modelos com dados desbalanceados. Por fim, para mitigar a falta de detalhes na explanação do estudo, deve-se utilizar guias que auxiliem a relatar de maneira transparente o desenvolvimento do modelo como o STARD (*Standards for Reporting of*

[1] https://www.nsctotal.com.br/colunistas/dagmara-spautz/estado-oficializa-criterio-que-da-prioridade-a-mais-jovens-e-saudaveis

[2] https://www.estado.rs.gov.br/mesmo-com-expansao-de-leitos-45-rodada-confirma-pressao-sobre-capacidade-hospitalar-e-rs-em-bandeira-preta

Diagnostic Accuracy) (BOSSUYT et al., 2003) e o TRIPOD (*Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis*) (COLLINS et al., 2015).

Neste trabalho, propomos um modelo preditivo baseado em florestas aleatórias (BREIMAN, 2001) para classificação de risco de óbito para pacientes confirmados de COVID-19 no estado do Rio Grande do Sul/Brasil. Utilizamos a pontuação AUC-ROC para avaliar a performance do modelo. Este é um indicador importante pois nos fornece uma medida da precisão total do modelo independente de um limiar particular (FAWCETT, 2006). O desenvolvimento do modelo é explanado utilizando o guia TRIPOD para atenuar o risco da falta de detalhes. Utilizamos também da validação cruzada 5-*fold* para validar internamente o modelo e, desta forma, reduzir o risco de enviesamento.

O restante deste trabalho está organizado como segue. Na Seção 2 é apresentado o estado da arte do aprendizado de máquina aplicado ao combate da pandemia. Na Seção 3 é discutido o projeto do estudo, a aquisição dos dados e a construção do modelo de prognóstico. Na Seção 4 avaliamos a performance do modelo construído, bem como a sua interpretabilidade. Por fim, na Seção 5 há a discussão geral sobre as contribuições e limites do presente estudo.

2 ESTADO DA ARTE

Nesta seção é apresentado o estado da arte do aprendizado de máquina aplicado ao combate da pandemia. Aprendizagem de Máquina é um ramo da Inteligência Artificial onde sistemas computacionais podem aprender a partir de dados e identificar padrões com o mínimo de intervenção humana (MITCHELL et al., 1997). A Tabela 1 apresenta o resumo de métodos de aprendizagem de máquina, métricas de avaliação e características aplicadas ao contexto da pandemia de COVID-19. Os métodos identificados tem por objetivo auxiliar na tomada de decisão médica, incluindo prever a severidade dos casos confirmados (classificação de risco), chance de óbito ou recuperação, e chance de internação em UTI.

Segundo os dados coletados, o método Floresta Aleatória é o mais frequentemente utilizado (aparece em 80% dos casos). Isto ocorre devivo ao fato deste método ser considerado acurado e robusto pois conta com um alto número de árvores de decisão no processo de treinamento, por não sofrer com sobreajuste e pelo seu auxílio na classificação da importância das características (BREIMAN, 2001). Devido às características mencionadas, utilizamos este algoritmo no processo de desenvolvimento do modelo preditivo para classificação de risco de mortalidade de pacientes com COVID-19.

As principais características utilizadas para treinamento são dados demográficos (*e.g.*, idade, sexo e país de origem), presentes em 50% dos trabalhos; dados clínicos (*e.g.*, sintomas e comorbidades), presentes em 100% dos trabalhos; dados laboratoriais (*e.g.*, amostras de sangue), presentes em 40% trabalhos, onde foi possível identificar a mínima saturação de O2 como característica chave. Por fim, características radiológicas (*e.g.*, imagens de radiografia) também aparecem como características de treinamento (CASIRAGHI et al., 2020).

Nos modelos propostos e treinados, a quantidade de características usadas varia de 3 (YADAW et al., 2020) a 99 (CHENG et al., 2020). Entretanto, foi possível observar que a idade e o sexo foram unânimes nos estudos que utilizaram de dados demográficos. Além disso, é importante ressaltar que nos modelos que utilizam dados clínicos, em termos de sintomas e comorbidades, todos utilizaram características relacionadas a resfriados, como febre, tosse e dispnéia.

Para entender quão bom um modelo preditivo é, existem métricas para avaliar o desempenho (ou generalização) de um método de aprendizagem de máquina (AMIDI; AMIDI, 2020). As métricas mais encontradas nos trabalhos foram métricas obtidas através da matriz de confusão, sendo a AUC-ROC a mais utilizada (presente em 80% dos trabalhos), seguida por acurácia, precisão, sensibilidade, especificidade e F1 *score* (pre-

Tabela 1 – Métodos de aprendizagem de máquina, métricas de avaliação e características

	Métodos de aprendizagem de máquina usados	**Características**	**Métricas de avaliação**
(IWENDI et al., 2020)	Floresta Aleatória* Intensificada (AdaBoost); Árvore de decisão; Máquina de Vetores de Suporte (SVM); Naive Bayes Gaussiano	Demográficas; Clínicas	Acurácia; Precisão; Revocação; F1 Score; Matriz de confusão
(POURHOMAYOUN; SHAKIBI, 2020)	Redes Neurais Artificiais; Árvore de decisão; Floresta Aleatória*; K-vizinhos mais próximos (KNN); Regressão Logística; Máquina de Vetores de Suporte (SVM)	re Demográficas; Clínicas	Acurácia; ROC-AUC; Matriz de confusão
(ZHAO et al., 2020)	Regressão Logística	Clínicas	ROC-AUC
(YADAW et al., 2020)	eXtreme Gradient Boosting (XGBoost); Regressão Logística; Floresta Aleatória*; Máquina de Vetores de Suporte (SVM)	Demográficas; Clínicas	ROC-AUC
(CHENG et al., 2020)	Floresta Aleatória*	Clínicas; Laboratoriais	Revocação; Especificidade; Acurácia; ROC-AUC; Matriz de confusão
(ASSAF et al., 2020)	Redes Neurais Artificiais; Floresta Aleatória*; Árvore de Decisão	Clínicas	Revocação; Especificidade; Acurácia; F1 Score; ROC-AUC; Matriz de confusão
(GAO et al., 2020)	Regressão Logística; Máquina de Vetores de Suporte (SVM); eXtreme Gradient Boosting (XGBoost); Redes Neurais Artificiais; K-vizinhos mais próximos (KNN)	Clínicas; Laboratoriais	Razão de probabiliadde positiva; Razão de probabilidade negativa; F1 score; ROC-AUC; Matriz de confusão
(CHOWDHURY et al., 2020)	eXtreme Gradient Boosting (XGBoost); Regressão Logística	Demográficas; Clínicas; Laboratoriais	Revocação; Especificidade; Razão de probabiliadde positiva; Razão de probabilidade negativa; ROC-AUC; Matriz de confusão
(CASIRAGHI et al., 2020)	Floresta Aleatória*	Radiológicas; Clínicas; Laboratoriais	Revocação; Especificidade; Acurácia; F1 score; ROC-AUC; Matriz de confusão
(DUN et al., 2020)	Floresta Aleatória*	Demográficas; Clínicas	–

sentes em 70% dos trabalhos).

Os estudos mostram que modelos de aprendizagem de máquina conseguem prever chances de recuperação ou óbito com até 94% de acurácia, com um F1 *score* de 0.84 (IWENDI et al., 2020). Essa métrica (F1 *score*) auxilia na verificação do quão bom e generalista o modelo é, visto que é uma média harmônica entre a precisão (verdadeiros

que realmente eram verdadeiros) e a sensibilidade (proporção dos verdadeiros positivos entre todas as observações que realmente são positivas). Isto demonstra a importância de outras métricas para avaliação dos modelos, além da acurácia, que é tipicamente a métrica mais utilizada em problemas de classificação binária.

Os trabalhos relacionados na Tabela 1 possuem uma limitação em comum: o grupo de pacientes utilizados para treino. Um modelo de previsão aplicado em um novo sistema de saúde, configuração ou país geralmente produz previsões que estão mal calibradas e podem precisar ser atualizadas antes que possam ser aplicadas com segurança nessa nova configuração. Os mesmos métodos, com outras populações ou de outros sistemas de saúde, podem levar a resultados diferentes (ASSAF et al., 2020; CHENG et al., 2020; ZHAO et al., 2020; CASIRAGHI et al., 2020).

Dados de vários países e sistemas de saúde podem permitir uma melhor compreensão da generalização e implementação de modelos de previsão em diferentes configurações e populações (WYNANTS et al., 2020). É com esta premissa que neste trabalho serão utilizados dados abertos, as mesmas técnicas de aprendizagem de máquina, avaliação e validação do estado da arte, porém aplicadas à população do Rio Grande do Sul.

3 METODOLOGIA

3.1 Projeto do Estudo

Objetivando mitigar problemas relacionados ao risco de enviesamento por pobreza no relatório sobre o desenvolvimento do modelo (WYNANTS et al., 2020), a metodologia para o desenvolvimento e exposição dos resultados seguem o guia TRIPOD (COLLINS et al., 2015). Este é um *checklist* sobre pontos essenciais que um estudo de diagnóstico ou prognóstico deve relatar quando estiver explanando o desenvolvimento de um modelo preditivo multivariado. O *checklist* TRIPOD deste trabalho pode ser encontrado no apêndice A.

3.2 Participantes

Para o desenvolvimento, treino e validação do modelo preditivo, utilizamos dados anonimizados provenientes do Painel Coronavírus RS[3]. A escolha dos dados foi devido a acessibilidade dos mesmos e também por estarem isentos das leis de proteção aos dados (MACHADO et al., 2019). Neste estudo, utilizamos um conjunto de dados com 604.389 registros (observações) de casos confirmados de COVID-19. O conjunto de dados original contém um total de 30 variáveis em nível-paciente, incluindo dados demográficos (*e.g.*, sexo e faixa-etária), clínicos (*e.g.*, tosse e dispneia) e temporais (*e.g.*, data do início dos sintomas), compreende o perído de 01 de janeiro de 2021 à 08 de junho de 2021.

3.2.1 Engenharia de características

A Figura 1 retrata o fluxo da seleção dos dados elegíveis para este estudo. No conjunto de dados, existem registros de pacientes que vieram a óbito por causas distindas da COVID-19 (20 [$< 0{,}01\%$]), e também há aqueles cujo parecer evolutivo ainda não era conhecido naquele momento, ou seja, que ainda estavam em acompanhamento (17.724 [$< 3\%$]). Há também registros com informação faltante do sintoma dispneia (2.418 [$<$ 0,5]%). Optamos por remover os registros com informações faltantes e aqueles que não possuem informação relevante para o estudo (*i.e.*, óbitos por outras causas ou em acompanhamento). Ao final da seleção, dos 604.389 registros inicais, 584.228 permaneceram na base para prosseguir o estudo.

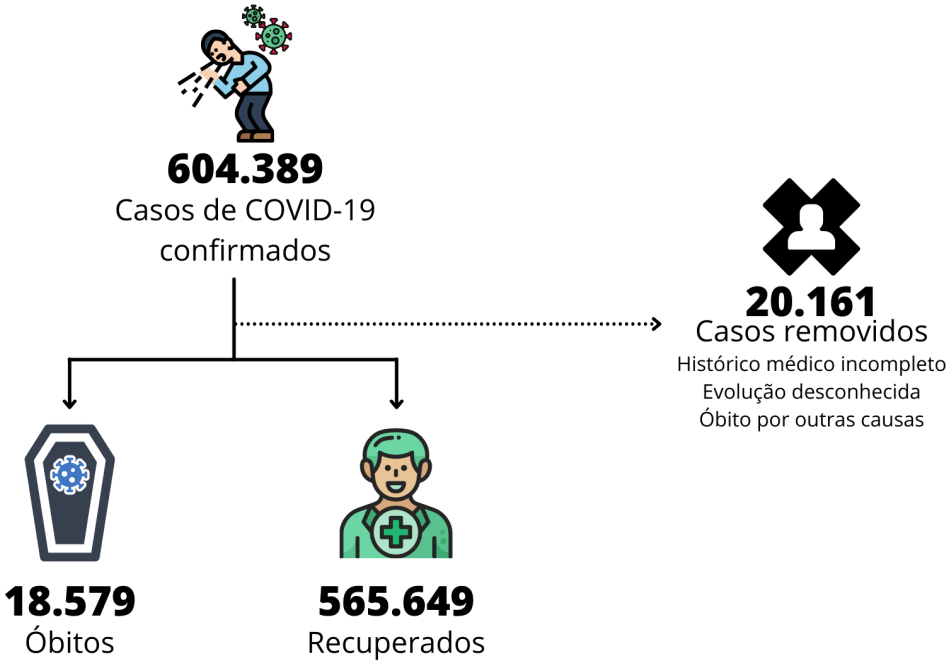
Além das características iniciais do conjunto de dados[4], outras seis características

[3] https://ti.saude.rs.gov.br/covid19/
[4] https://ti.saude.rs.gov.br/covid19/api

foram extraídas do campo textual "CONDICOES" (campo pré-existente no conjunto de dados), a saber: cardiopatia, diabetes, doença respiratória, problema renal, obesidade e doença cromossômica. Como a variável "Raça/cor" possuía 19,98% (n=116.752) de dados faltantes, sendo estes 58% dos óbitos totais (n=10.889), e devido ao decréscimo da pontuação do modelo ao adicioná-la como preditora (ver Seção 4), optamos por sua remoção do conjunto de dados.

Utilizamos da codificação por rótulo (*label encoding)* nas variáveis pois algoritmos de aprendizagem de máquina performam algebra linear no processo de treinamento (CHATFIELD et al., 2011). Como a distribuição das preditoras é binária, transformamos "SIM" para 1 e "NÃO" para 0. A única exceção no processo é da variável qualitativa ordinal "FAIXAETARIA". Neste caso, cada faixa-etária foi codificada com um número, a saber: "< 1": 0; "01 a 04": 1; "05 a 09": 2; "10 a 14": 3; "15 a 19": 4; "20 a 29": 5, "30 a 39": 6; "40 a 49": 7; "50 a 59": 8; "60 a 69": 9; "70 a 79": 10; "80 e mais": 11.

Figura 1 – Fluxo de seleção dos participantes



3.3 Variável resposta

O principal objetivo do trabalho é prever a chance de óbito entre pacientes positivos de COVID-19. Esta análise prognóstica é importante para a priorização dos pacientes na alocação dos recursos hospitalares. Os dados coletados possuem a informação relacionada a evolução do paciente na variável "EVOLUCAO". Esta variável é dicotômica – "ÓBITO" (1) se o paciente faleceu, e "RECUPERADO" (0) caso contrário – e nos permite a criação de um modelo de aprendizagem de máquina para classificar os pacientes quanto à sua

chance de falecimento, pois informa quais e quantos foram recuperados ou vieram a óbito por COVID-19.

3.4 Variáveis Preditoras

As variáveis candidatas para preditoras foram selecionadas baseando-se em um dos critérios comuns (COLLINS et al., 2015): variáveis clínicas conhecidas de diagnóstico comum relacionado a pneumonia ou quadros gripais. Após a extração das novas características e remoção de campos que não são candidatos a variáveis preditoras (*i.e.*, datas e códigos), os dados clínicos, demográficos e categóricos resultantes e que foram utilizados no estudo podem ser vistos na Tabela 2.

Tabela 2 – Variáveis

Nome	Descrição	Tipo
SEXO	Sexo	Binária/Demográfica
FAIXAETARIA	Faixa etária	Categórica/Demográfica
EVOLUCAO	Qual o parecer evolutivo do paciente?	Binária/Alvo
FEBRE	Sintomas de febre	Binária/Clínica
TOSSE	Sintomas de tosse	Binária/Clínica
GARGANTA	Sintomas de dor de garganta	Binária/Clínica
DISPNEIA	Sintomas de dispnéia/ falta de ar	Binária/Clínica
GESTANTE	O paciente é gestante?	Binária/Clínica
SRAG	O quadro clínico se caracteriza como Sindrome Respiratória Aguda Grave?	Binária/Clínica
CARDIOPATIA	O paciente possui alguma doença cardíaca crônica?	Binária/Clínica
DIABETES	O paciente possui diabetes?	Binária/Clínica
DOENCA_RESPIRATORIA	O paciente possui alguma doença respiratória crônica?	Binária/Clínica
PROBLEMA_RENAL	O paciente possui problema renal?	Binária/Clínica
OBESIDADE	O paciente é obeso?	Binária/Clínica
DOENCA _CROMOSSOMICA	O paciente possui doença cromossômica?	Binária/Clínica

3.5 Métodos Estatísticos

3.5.1 Análise descritiva

Para entendermos sobre possíveis características que expliquem o risco de óbito, performamos análises descritivas dos preditores por cada grupo e apresentamos os re-

sultados em quantidade e proporção de óbitos. Possíveis correlações foram analisadas utilizando o coeficiente de Pearson (BENESTY et al., 2009) e a importância relativa das características foi analisada através do grau de impureza de Gini (MENZE et al., 2009) (*i.e.*, a importância da característica para o modelo).

3.5.2 Análise preditiva

Aplicamos um algoritmo de aprendizagem de máquina para prever a chance de óbito entre pacientes confirmados com COVID-19. Para a implementação, análise estatística, avaliação e validação dos modelos, utilizamos a linguagem Python 3, juntamente com as bibliotecas para aprendizagem de máquina: Scikit-learn (PEDREGOSA et al., 2011) (floresta aleatória), e de manipulação e visualização de dados Pandas (MCKINNEY, 2010), Numpy (HARRIS et al., 2020) e Seaborn (HUNTER, 2007).

Para garantir a reprodutibilidade do experimento, definimos arbritrariamente a semente 777 para embaralhar o conjunto de dados inicial. Utilizamos uma divisão pseudo-aleatória (*random split*) de 70%/30%, a partir dos dados iniciais, sendo 70% utilizado para treinos e testes, e os 30% restante, para validação interna do modelo (JAMES et al., 2013). Avalimamos a performance/discriminação do modelo através da área sobre a curva ROC (AUC-ROC). A pontuação AUC-ROC pode assumir valores de 0 a 1, onde um valor de 0 indica um modelo que erra todas as predições e 1 reflete um teste perfeito de predição. Em geral, um valor de 0,5 sugere que o modelo não tem capacidade de discriminação (*i.e.*, não consegue diferenciar pacientes que vieram a óbito dos recuperados); 0,7 a 0,8 é considerado aceitável; 0,8 a 0,9 é ótimo; e valores acima de 0,9 são considerados excelentes (MANDREKAR, 2010).

Por se tratar de um conjunto de dados desbalanceado, utilizamos da técnica de subamostragem (*under-sampling*, do inglês) para balancear os dados de treino, e então confrontar com o treinamento sem balanceamento para detectar melhorias (ANAND et al., 2010). Esta técnica consiste em selecionar, pseudo-aleatóriamente, pequenas fatias da classe majoritária (recuperado), a fim de igualar à quantidade de dados pertencentes a classe minoritária (óbito). Utilizamos da pesquisa de grade por validação cruzada (*gridsearch cross validation*) para encontrar os melhores hiperparâmetros para o modelo - incluindo a quantidade de árvores de decisão (`n_estimators`) e a profundidade máxima de cada árvore (`max_depth`) (KRSTAJIC et al., 2014) - que maximizassem a pontuação AUC-ROC nos dados de treino.

3.5.3 Validação do modelo

Para validação interna do modelo, utilizamos da técnica *K-Fold Cross Validation*, que consiste em dividir o conjunto de dados em K dobras (*folds*). A função de previsão

é aprendida usando K-1 dobras, e a dobra deixada de fora é usada para teste. Optamos por utilizar sua variação estratificada (*stratified K-Fold Cross Validation*), ou seja, na adaptação para conjuntos de dados desbalanceados, pois mantém a proporção original de cada classe na etapa de teste. Geralmente, realiza-se a validação cruzada k-fold usando k = 5 ou k = 10, uma vez que estes valores produzem estimativas de taxa de erro de teste que não sofrem de viés ou variância excessivamente altos (JAMES et al., 2013). Por padrão, a biblioteca Scikit-learn (PEDREGOSA et al., 2011) utiliza K = 5, e por isso decidimos manter este valor.

4 RESULTADOS

4.1 Participantes

O perfil dos participantes é apresentado na Tabela 3. Do conjunto de dados utilizado no estudo, 565.649 (96,82%) casos possuem evolução conhecida como recuperação, e 18.579 (3,18%) são casos que tiveram óbito comprovado por COVID-19, onde a taxa de letalidade geral é de 3,07% (n=18.579). Dos casos confirmados, pouco mais da metade é do sexo feminino (53,10%, n=310.263), tendo o sexo masculino a maior taxa de letalidade proporcional (3,6%, n=9.889). Entre as faixa-etárias, a maior concentração está entre 30 e 39 anos (20,86%), sendo a faixa superior aos 70 anos a mais letal (43.74%, n=7.909). Dentre os sintomas, destacamos a falta de ar/dispnéia como o sintoma com a maior taxa de letalidade (17,49%, n=15.676). Em contra-ponto, a dor-de-garganta é o sintoma apresentado que possui a menor taxa de letalidade (1,39%, n=2.952).

Tabela 3 – Dados dos participantes

Nome do atributo		Infectados	Óbitos	Proporção (%)
SEXO	Masculino	273.965	9889	3,60
	Feminino	310.263	8.690	2,80
FAIXA_ETARIA	<1	4.133	4	0,09
	01 a 04	7.403	4	0,05
	05 a 09	9.267	2	0,02
	10 a 14	13390	5	0,03
	15 a 19	30.292	19	0,06
	20 a 29	108.501	225	0,20
	30 a 39	121.926	758	0,62
	40 a 49	104.102	1.716	1,64
	50 a 59	88.424	3.138	3,54
	60 a 69	57.058	4.799	8,41
	70 a 79	27.233	4.516	16,58
	80 e mais	12.490	3.393	27,16
FEBRE	Sim	198.357	8.909	4,49
	Não	385.871	9.670	2,50
TOSSE	Sim	270.337	10.873	4,02
	Não	313.891	7.706	2,45
GARGANTA	Sim	211.297	2.952	1,39
	Não	372.931	2.952	4,19

DISPNEIA	Sim	89.606	15.676	17,49
	Não	494.622	2.903	0,58
GESTANTE	Sim	3.157	47	1,48
	Não	581.071	18.532	3,18
SRAG	Sim	51.691	18.579	35,94
	Não	532.537	0	0,0
CARDIOPATIA	Sim	16.981	7.728	45,50
	Não	556.396	10.851	1,91
DIABETES	Sim	22.627	5.593	24,71
	Não	561.601	17.034	2,31
DOENCA_RESPIRATORIA	Sim	3.566	1.552	43,52
	Não	563.635	17.027	2,93
PROBLEMA_RENAL	Sim	2254	861	38,19
	Não	581.974	17.718	3,04
OBESIDADE	Sim	10.855	2.868	26,42
	Não	573.373	15.711	2,74
DOENCA_CROMOSSOMICA	Sim	141	83	41,13
	Não	584.087	18.521	3,17

4.2 Desenvolvimento do modelo

Aplicamos o algoritmo de Floresta aleatória, utilizando 100 árvores de decisão (`n_estimators`) no processo e com 4 de profundidade máxima (`max_depth`), pois foram os melhores hiperparâmetros retornados pela busca em grade com validação cruzada (k = 5). Como o conjunto de dados de teste é desbalanceado, possuindo 395.954 (96,82%) casos recuperados e 13.005 (3,18%) óbitos, optamos por utilizar da técnica de subamostragem aleatória (*random under sampling*) (ANAND et al., 2010) nos dados de treino.

O coeficiente de correlação de Pearson entre as variáveis independentes pode ser observado na Figura 2. Podemos constatar uma variação de -0,8 a 0,6, e que as variáveis que possuem uma maior dependência (>= 0.2) com a variável resposta (EVOLUCAO) são SRAG, DISPNEIA, CARDIOPATIA, DIABETES, DOENCA_RESPIRATORIA e OBESIDADE (*i.e.*, variáveis que possuem maior correlação). E de acordo com grau de impureza de Gini, constatamos a influência relativa das *top* 5 características na classificação predita pelo modelo, a saber: SRAG, FAIXAETARIA, CARDIOPATIA, DISPNEIA e DIABETES (Figura 3).

Figura 2 – Matriz de correlação de Pearson

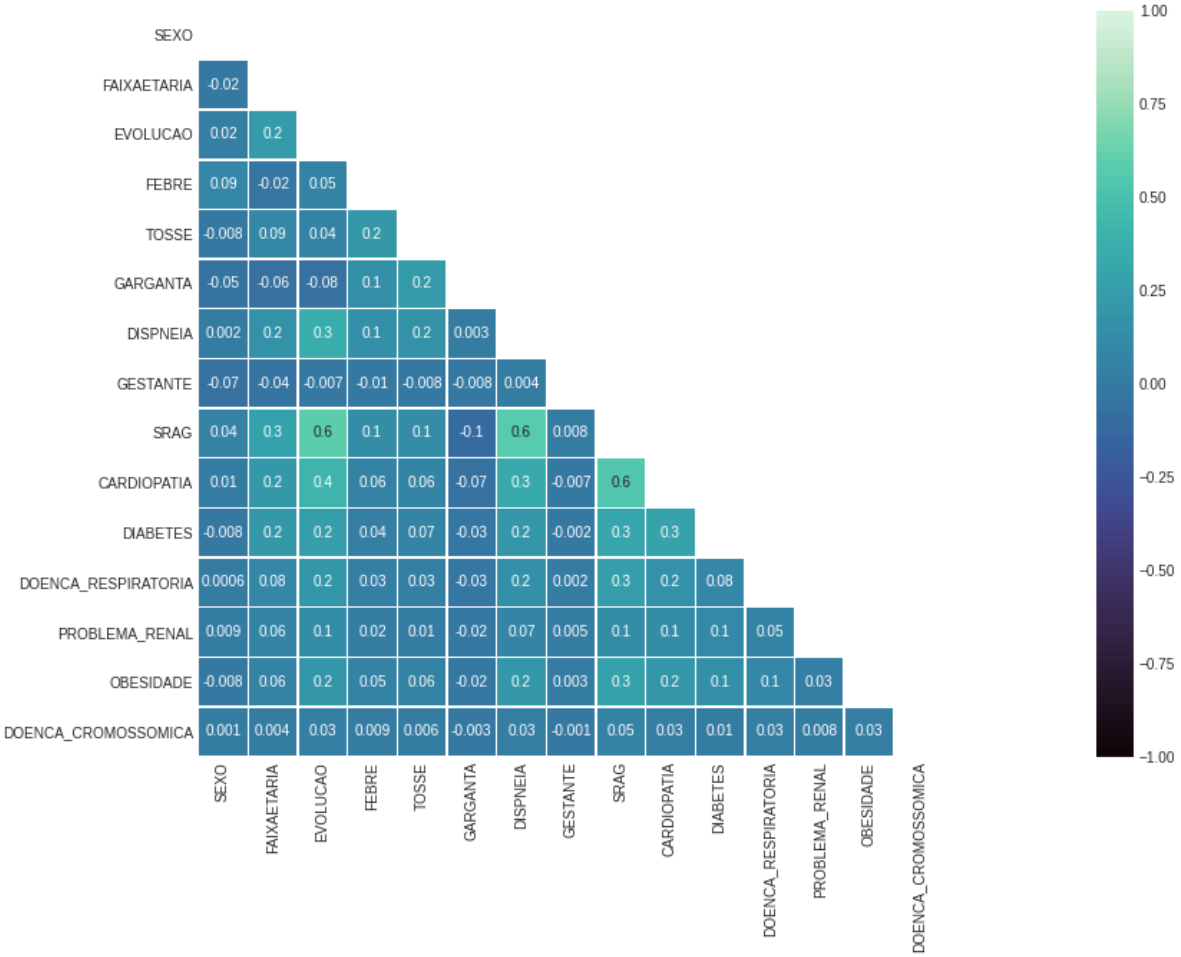
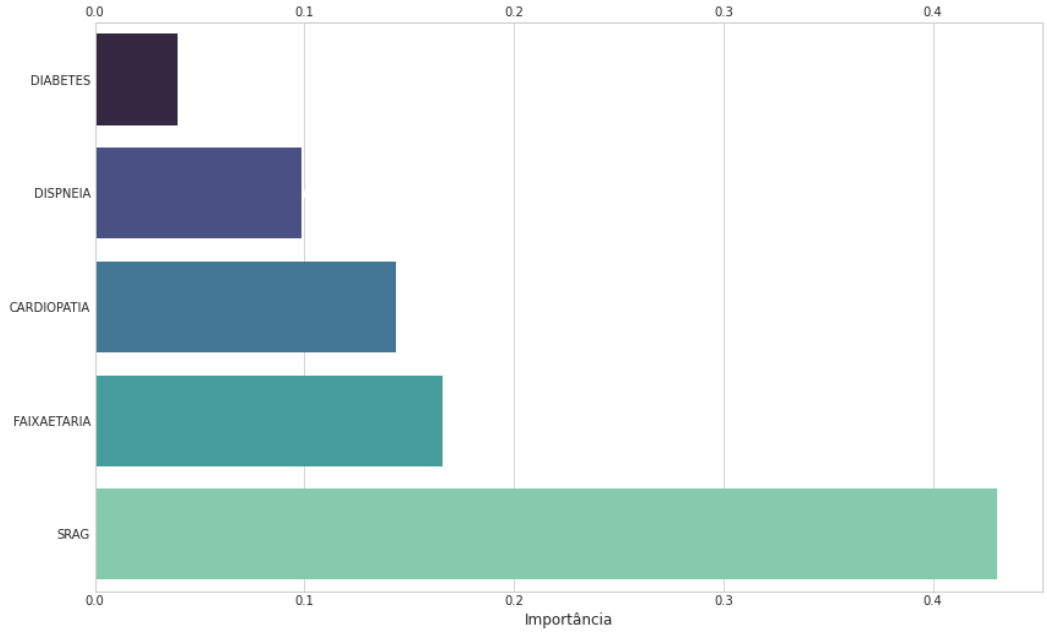
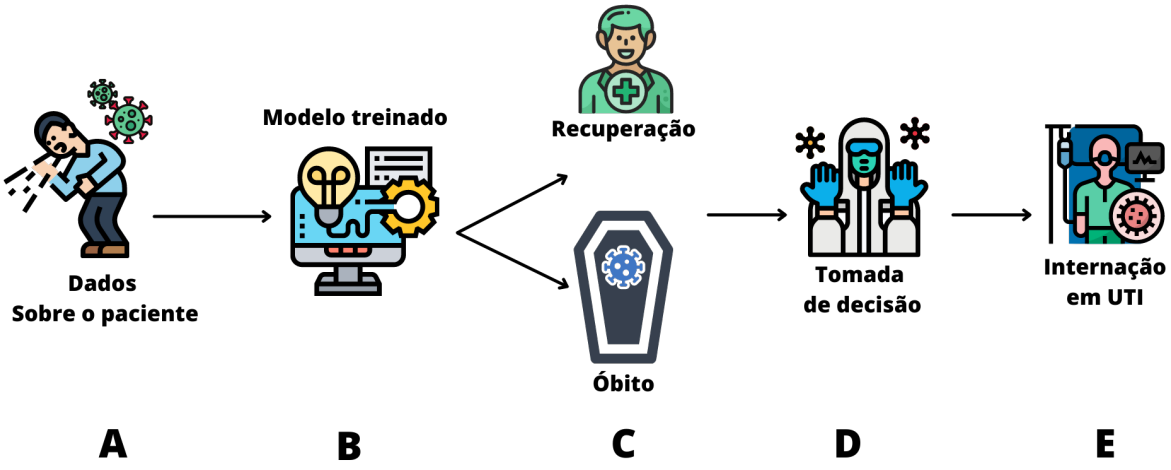


Figura 3 – Índice de impureza de Gini



O fluxo geral para o uso do modelo proposto pode ser visualizado na Figura 4. Primeiramente, os dados clínicos e sintomas do paciente são coletados (A) e inseridos no modelo já treinado (B). Com o desfecho predito pelo modelo (C), o profissional da saúde em atividade (D) pode basear-se para tomar a decisão se aquele indivíduo pode ou não ocupar a vaga de leito de UTI (E).

Figura 4 – Fluxo de uso do modelo



4.3 Avaliação e Validação do modelo

O conjunto inicial dos dados possuía 584.228 registros e um total de 18.579 óbitos. O conjunto de treino e teste possui 408.959 registros, sendo 3,18% (n=13.005) óbitos. O conjunto de validação possui a mesma distribuição, sendo 5.574 óbitos. A Figura 5 ilustra a divisão dos dados em cada etapa de treino, teste e validação. Os valores das pontuações médias após a validação cruzada (5-V.C.) do modelo de floresta aleatória, juntamente com seus intervalos de confiança com nível de confiança de 95% (I.C. = 95%) podem ser vistos na Tabela 4. O modelo com menos registros de óbito (com raça/cor) possui o pior desempenho e foi descartado das melhorias posteriores (subamostragem e ajuste de hiperparâmetros).

Figura 5 – Fluxo dos dados de treino, teste e validação

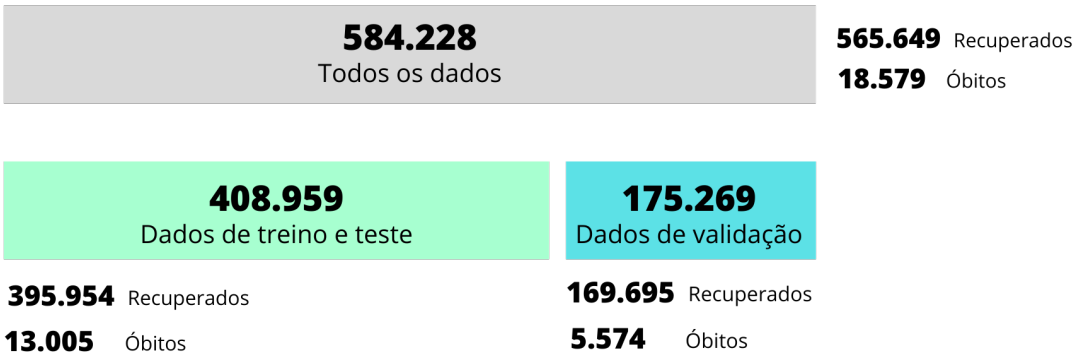
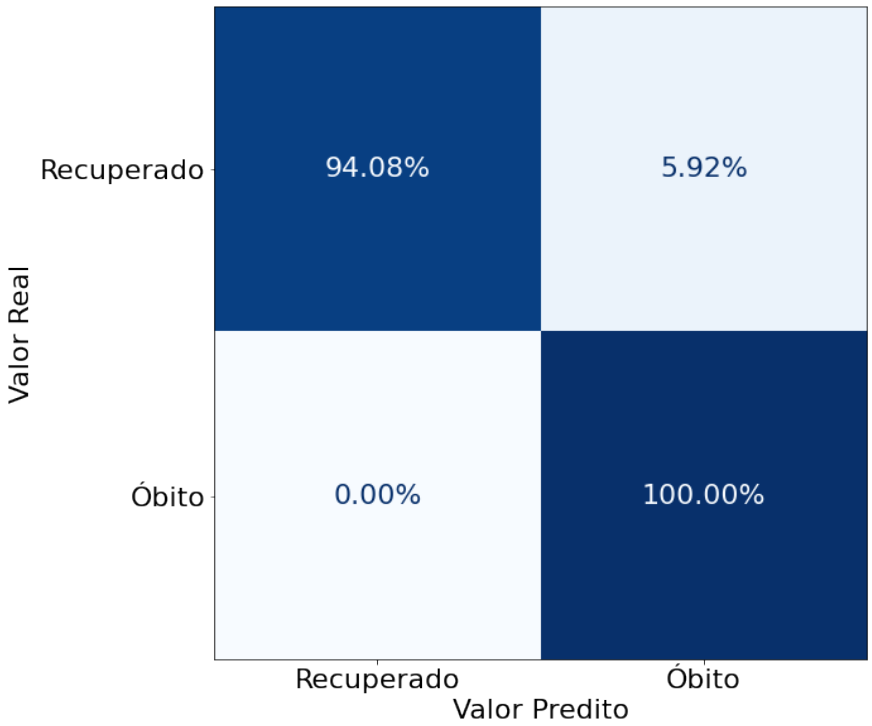


Tabela 4 – Performance do modelo

	Pontuação AUC-ROC	
	Média (5-V.C.)	I.C. = 95 %
Floresta aleatória (com raça/cor)	0,539	(0,528 - 0,550)
Floresta aleatória (linha de base)	0,705	(0,694 - 0,716)
Floresta aleatória (subamostragem aleatória)	0,969	(0,968 - 0,970)
Floresta aleatória (subamostragem aleatória + ajuste de hiperparâmetro)	0,981	(0,981 - 0,982)

Utilizamos uma matriz de confusão (Figura 6) para descrever e visualizar a performance do classificador nos dados de validação e também para prover *insights* sobre onde o modelo errou. É importante salientar que pontuação AUC-ROC é derivada desta matriz, que informa os falsos positivos e negativos, bem como os classificados corretamente. No nosso contexto, a nossa classe positiva são os casos de óbitos e a classe negativa são os casos recuperados. Podemos observar na matriz que apenas 5,92% (n=10.045) casos recuperados foram classificados errôneamente (falsos positivos) e que todos os casos de óbito foram classificados corretamente (*i.e.*, sem falsos negativos). Como o modelo performou acima de 0,90 de pontuação AUC-ROC, podemos classificá-lo como excelente discriminador (JR; LEMESHOW; STURDIVANT, 2013).

Figura 6 – Matriz de confusão conjunto de validação



5 DISCUSSÃO

5.1 Limitações

Este estudo possui pelo menos quatro limitações. Primeiramente, alguns dos maiores preditores laboratoriais utilizados na construção de modelos prognósticos (*i.e.*, proteína C-reativa e nível de linfócitos no sangue (WYNANTS et al., 2020)) não estão disponíveis no Painel Coronavírus RS, e por isso fomos impossibilitados de utilizar neste trabalho. Acreditamos que elas podem ajudar a construir um modelo com maior poder preditivo.

Segundo, a metodologia de treinamento do modelo adotada é denominada "aprendizado *offiline*", onde o nosso conjunto de dados é estático, ou seja, a quantidade de observações corresponde a um intervalo de tempo bem definido. Para mitigar o problema relacionado a possíveis mudanças devido a novas cepas e mutações do vírus, técnicas como o "aprendizado *online*" podem ser abordadas e comparadas com a solução do presente trabalho. Esta outra abordagem de aprendizado consiste em treinar o modelo à medida em que novos dados são gerados.

É importante salientar também, que analisamos os dados como sendo registros de pacientes únicos. Isto ocorre devido aos dados de reinfecção não estarem disponíveis no Painel Coronavírus RS e implica que eventuais casos de reinfecção foram ignorados. E por fim, como utilizamos apenas um algoritmo de aprendizagem de máquina no processo, em pesquisas futuras outros métodos de análise preditiva podem ser utilizados para comparar com o desempenho do algoritmo de florestas aleatórias.

5.2 Conclusões e implicações políticas

Desenvolvemos e validamos internamento um classificador que permite estratificar com precisão o risco de óbito de pacientes com COVID-19. O modelo preditivo proposto pode contribuir para a diminuição da subjetividade na tomada de decisão, para que o profissional de saúde possa tomar suas decisões baseadas não somente no empirismo, mas também em um parâmetro técnico dentro do ambiente hospitalar. O código e os dados estão disponíveis no repositório no GitHub: <https://github.com/gustavocrod/predict_death_covid>. Entretanto, salientamos que os aspectos éticos e legais devem ser considerados no momento em que os profissionais tomam a decisão, ou seja, esta é apenas uma ferramenta adicional que pode auxiliar no processo de tomada de decisão e não substitui a expertise médica.

REFERÊNCIAS

AMIDI, A.; AMIDI, S. **Machine Learning tips and tricks cheatsheet**. [S.l.], 2020. Disponível em: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>. Citado na página 21.

ANAND, A. et al. An approach for classification of highly imbalanced data using weighting and undersampling. **Amino acids**, Springer, v. 39, n. 5, p. 1385–1391, 2010. Citado 2 vezes nas páginas 28 e 32.

ASSAF, D. et al. Utilization of machine-learning models to accurately predict the risk for critical covid-19. **Internal and emergency medicine**, Springer, v. 15, n. 8, p. 1435–1443, 2020. Citado 3 vezes nas páginas 19, 22 e 23.

BENESTY, J. et al. Pearson correlation coefficient. In: **Noise reduction in speech processing**. [S.l.]: Springer, 2009. p. 1–4. Citado na página 28.

BOSSUYT, P. M. et al. The stard statement for reporting studies of diagnostic accuracy: explanation and elaboration. **Annals of internal medicine**, American College of Physicians, v. 138, n. 1, p. W1–12, 2003. Citado na página 20.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 20 e 21.

CASIRAGHI, E. et al. Explainable machine learning for early assessment of covid-19 risk prediction in emergency departments. **IEEE Access**, IEEE, v. 8, p. 196299–196325, 2020. Citado 4 vezes nas páginas 19, 21, 22 e 23.

CHATFIELD, K. et al. The devil is in the details: an evaluation of recent feature encoding methods. In: **BMVC**. [S.l.: s.n.], 2011. v. 2, n. 4, p. 8. Citado na página 26.

CHENG, F.-Y. et al. Using machine learning to predict icu transfer in hospitalized covid-19 patients. **Journal of Clinical Medicine**, Multidisciplinary Digital Publishing Institute, v. 9, n. 6, p. 1668, 2020. Citado 4 vezes nas páginas 19, 21, 22 e 23.

CHOWDHURY, M. E. et al. An early warning tool for predicting mortality risk of covid-19 patients using machine learning. **arXiv preprint arXiv:2007.15559**, 2020. Citado na página 22.

COLLINS, G. S. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. **Circulation**, Am Heart Assoc, v. 131, n. 2, p. 211–219, 2015. Citado 3 vezes nas páginas 20, 25 e 27.

DUN, C. et al. A machine learning study of 534,023 medicare beneficiaries with covid-19: Implications for personalized risk prediction. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Citado na página 22.

FAWCETT, T. An introduction to roc analysis. **Pattern recognition letters**, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 20.

GAO, Y. et al. Machine learning based early warning system enables accurate mortality risk prediction for covid-19. **Nature communications**, Nature Publishing Group, v. 11, n. 1, p. 1–10, 2020. Citado na página 22.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, v. 585, p. 357–362, 2020. Citado na página 28.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in science & engineering**, IEEE Computer Society, v. 9, n. 3, p. 90–95, 2007. Citado na página 28.

IWENDI, C. et al. Covid-19 patient health prediction using boosted random forest algorithm. **Frontiers in public health**, Frontiers, v. 8, p. 357, 2020. Citado na página 22.

JAMES, G. et al. **An introduction to statistical learning**. [S.l.]: Springer, 2013. v. 112. Citado 2 vezes nas páginas 28 e 29.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. [S.l.]: John Wiley & Sons, 2013. v. 398. Citado na página 35.

KRSTAJIC, D. et al. Cross-validation pitfalls when selecting and assessing regression and classification models. **Journal of cheminformatics**, BioMed Central, v. 6, n. 1, p. 1–15, 2014. Citado na página 28.

LATIF, S. et al. Leveraging data science to combat covid-19: A comprehensive review. **IEEE Transactions on Artificial Intelligence**, IEEE, 2020. Citado na página 19.

MACHADO, R. et al. Vazamentos de dados: Histórico, impacto socioeconômico e as novas leis de proteção de dados. In: **Anais da XVII Escola Regional de Redes de Computadores**. Porto Alegre, RS, Brasil: SBC, 2019. p. 154–159. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/errc/article/view/9230>. Citado na página 25.

MANDREKAR, J. N. Receiver operating characteristic curve in diagnostic test assessment. **Journal of Thoracic Oncology**, Elsevier, v. 5, n. 9, p. 1315–1316, 2010. Citado na página 28.

MCKINNEY, W. Data structures for statistical computing in python. In: WALT, S. van der; MILLMAN, J. (Ed.). **Proceedings of the 9th Python in Science Conference**. [S.l.: s.n.], 2010. p. 51 – 56. Citado na página 28.

MENZE, B. H. et al. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. **BMC bioinformatics**, Springer, v. 10, n. 1, p. 1–16, 2009. Citado na página 28.

MITCHELL, T. M. et al. Machine learning. McGraw-hill New York, 1997. Citado na página 21.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. **the Journal of machine Learning research**, JMLR. org, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 28 e 29.

POURHOMAYOUN, M.; SHAKIBI, M. Predicting mortality risk in patients with covid-19 using artificial intelligence to help medical decision-making. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020. Citado na página 22.

RANNEY, M. L.; GRIFFETH, V.; JHA, A. K. Critical supply shortages—the need for ventilators and personal protective equipment during the covid-19 pandemic. **New England Journal of Medicine**, Mass Medical Soc, v. 382, n. 18, p. e41, 2020. Citado na página 19.

WYNANTS, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. **bmj**, British Medical Journal Publishing Group, v. 369, 2020. Citado 4 vezes nas páginas 19, 23, 25 e 37.

YADAW, A. S. et al. Clinical features of covid-19 mortality: development and validation of a clinical prediction model. **The Lancet Digital Health**, Elsevier, v. 2, n. 10, p. e516–e525, 2020. Citado 2 vezes nas páginas 21 e 22.

ZHAO, Z. et al. Prediction model and risk scores of icu admission and mortality in covid-19. **PloS one**, Public Library of Science San Francisco, CA USA, v. 15, n. 7, p. e0236618, 2020. Citado 3 vezes nas páginas 19, 22 e 23.

Apêndices

APÊNDICE A – TRIPOD *CHECKLIST*

SEÇÃO/TÓPICO	Item	Checklist Item	Página
Título e Resumo			
Título	1	Identificar se o estudo é o desenvolvimento e/ou validação de um modelo de predição multivariado, a população alvo, e o *desfecho* para ser predito	1
Resumo	2	Prover uma sumarização dos objetivos, *design* do estudo, configurações, participantes, tamanho da amostra, variáveis preditoras, variável resposta, análise estatística, resultados e conclusões	13
Introdução			
Contextualização e Objetivos	3a	Explicar o contexto médico (Incluindo se é diagnóstico ou prognóstico), e relacionar com o desenvolvimento ou validação do modelo preditivo multivariado, incluindo referência para modelos existentes	15
	3b	Especificar os objetivos, incluindo o que o estudo descreve, se é o desenvolvimento de um modelo preditivo, a validação de um modelo ou ambos.	16
Métodos/metodologia			
Origem dos dados	4a	Descrever o *design* do estudo ou a origem dos dados (e.g., teste randomizado, coorte, ou registro dos dados), separadamente dos conjuntos de desenvolvimento e validação, se aplicável.	21
	4b	Especifique as principais datas do estudo, incluindo o início da acumulação; fim da acumulação; e, se aplicável, fim do acompanhamento	21
Participantes	5a	Especifique os elementos-chave do cenário do estudo (por exemplo, atenção primária, atenção secundária, população em geral), incluindo o número e a localização dos centros.	21
	5b	Descrever o critério de inclusão (elegibilidade) dos pacientes	21/22
	5c	Dar detalhes do tratamento recebido, se for relevante.	Não se aplica
Outcome	6a	Definir claramente o desfecho (*outcome*) que será predito pelo modelo, incluindo como e quando avaliado	22
	6b	Relate quaisquer ações para avaliação cega do resultado a ser predito.	Não se aplica
Preditores	7a	Definir claramente todas as variáveis preditoras usadas no desenvolvimento ou validação do modelo preditivo multivariado, incluindo como e quando elas foram medidas	22/23
	7b	Reportar quaisquer ações para avaliação cega das preditoras para o *outcome* e outras preditoras	Não se aplica
Tamanho da amostra	8	Explique como chegou ao tamanho do estudo.	21/22
Dados faltantes	9	Descrever como os dados faltantes foram tratados (e.g., análise de caso completo, imputação simples, imputação múltipla) com detalhs de quaisquer método de imputação.	21
Métodos estatísticos	10a	Descrever como as variáveis preditoras foram tratadas na análise	22
	10b	Especificar o tipo do modelo, todos os processos para a construção do modelo (incluindo qualquer seleção de variável preditora), e métodos para validação interna	24
	10c	Especificar todas as métricas usadas para avaliar a performance do modelo, e se for relevante, para comparar com múltiplos modelos	24/25
Grupos de risco	11	Prover detalhes de como grupos de risco foram criados, se feito.	Não se aplica
Resultados			
Participantes	13a	Descreva o fluxo de participantes durante o estudo, incluindo o número de participantes com e sem o desfecho e, se aplicável, um resumo do tempo de acompanhamento. Um diagrama pode ser útil.	27/28
	13b	Descreva as características dos participantes (dados demográficos básicos, características clínicas, preditores disponíveis), incluindo o número de participantes com dados ausentes para preditores e variável resposta	27/28
Desenvolvimento do modelo	14a	Especificar o número de participantes com e sem desfecho (*outcome events*) em cada análise	30/31
	14b	Se feito, relatar a associação não ajustada entre cada preditor candidato e desfecho	Não se aplica
Especificação do modelo	15a	Apresente o modelo de predição completo para permitir predições para indivíduos (ou seja, todos os coeficientes de regressão e interceptação do modelo ou sobrevivência de linha de base em um determinado ponto de tempo).	Não se aplica
	15b	Explicar como usar o modelo preditivo	30
Performance do modelo	16	Relate as medidas de desempenho (com ICs) para o modelo de previsão.	31
Discussão			
Limitações	17	Discutir quaisquer limitações do estudo (como a amostragem não ser representativa, poucas observações por preditor, dados faltantes)	33
Interpretação	18	Dar uma interpretação geral dos resultados, considerando os objetivos, limitações, resultados de estudos similares e outras evidências relevantes	33

Implicações	19	Discutir o potencial clínico de uso do modelo e implicações em pesquisas futuras	33
Outras informações			
Informações suplementares	20	Fornecer informações sobre a disponibilidade de recursos complementares, como protocolo de estudo, calculadora da Web e conjuntos de dados	Não se aplica
Apoio financeiro	21	Informar a fundação de apoio e a posição dos apoiadores no presente estudo	Não se aplica