Universidade Estadual de Maringá
Departamento de Estatística
Programa de Pós-graduação em Bioestatística

HEVANS VINICIUS PEREIRA

Estudo de Preditores da Evolução de Pacientes Internados por COVID-19

Maringá, PR
16 de outubro de 2024

HEVANS VINICIUS PEREIRA

Estudo de Preditores da Evolução de Pacientes Internados por COVID-19

Dissertação apresentada à Universidade Estadual de Maringá, como parte das exigências do Programa de Pós-Graduação em Bioestatística para a obtenção do título de Mestre.

Universidade Estadual de Maringá - UEM

Orientador: Prof. Dr. Brian Alvarez Ribeiro de Melo

Maringá, PR

16 de outubro de 2024

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

P436e

Pereira, Hevans Vinicius

Estudo de preditores da evolução de pacientes internados por COVID-19 / Hevans Vinicius Pereira. -- Maringá, PR, 2024.

39 f. : il. color., figs., tabs.

Orientador: Prof. Dr. Brian Alvarez Ribeiro de Melo.

Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Estatística, Programa de Pós-Graduação em Bioestatística, 2024.

1. COVID-19 - Probabilidade de mortalidade. 2. Inteligência Artificial. 3. Redes neurais. 4. COVID-19 - Fatores de risco. I. Melo, Brian Alvarez Ribeiro de, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Estatística. Programa de Pós-Graduação em Bioestatística. III. Título.

CDD 23.ed. 519.5

Jane Lessa Monção - CRB 9/1173

HEVANS VINICIUS PEREIRA

Estudo de Preditores da Evolução de Pacientes Internados por COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Bioestatística do Centro de Ciências Exatas da Universidade Estadual de Maringá, como requisito parcial para a obtenção do título de Mestre em Bioestatística.

BANCA EXAMINADORA

Prof. Dr. Brian Alvarez Ribeiro de Melo
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Willian Luis de Oliveira
Universidade Estadual de Maringá – PBE/UEM

Prof. Dr. Paulo César Ossani
Universidade Estadual de Maringá – DES/UEM

Maringá, 13 de setembro de 2024.

Dedico este trabalho a minha esposa Daiane, por todo o apoio.

Resumo

Este estudo teve como objetivo avaliar o desempenho de algoritmos de inteligência artificial (IA), mais especificamente regressão logística, floresta aleatória e redes neurais, na classificação de pacientes mais propensos a vir a óbito por COVID-19 no Brasil em 2022, com base em características clínicas, demográficas e hospitalares de cada paciente. Para tanto, utilizou-se um conjunto de dados amplo, extraído do Sistema de Informação de Vigilância Epidemiológica da Gripe, que registra todos os casos de internação por Síndrome Respiratória Aguda Grave (SRAG), incluindo infecções por COVID-19, no Brasil. Os algoritmos foram aplicados com o objetivo de identificar padrões e prever a probabilidade de mortalidade entre os pacientes hospitalizados com COVID-19. Além de prever a mortalidade, o estudo também buscou identificar as variáveis mais relevantes para o desfecho de óbito por COVID-19. Utilizou-se, para isso, o cálculo de importância das variáveis por meio de maneira intrísecas a cada algoritmo e da técnica SHAP (*SHapley Additive Explanations*), facilitando a interpretação dos modelos. Os algoritmos testados apresentaram bons níveis de desempenho alcançando a acurácia de cerca de 74% e AUC-ROC de 74%. Os resultados confirmaram que pacientes idosos, pacientes que fizeram uso de suporte ventilatório ou que necessitaram de internação em UTI estavam significativamente mais propensos a vir a óbito. Esses achados são consistentes com a literatura e reforçam a necessidade de intervenções precoces e direcionadas para esses grupos de risco.

Palavras-chave: COVID-19, Inteligência Artificial, Aprendizado de Máquina, Fatores de Risco, Regressão Logística, Floresta Aleatória, Redes Neurais.

Abstract

This study aimed to evaluate the performance of artificial intelligence (AI) algorithms, specifically logistic regression, random forest, and neural networks, in classifying patients most likely to die from COVID-19 in Brazil in 2022, based on each patient's clinical, demographic, and hospital characteristics. For this purpose, a large dataset was used, extracted from the Influenza Epidemiological Surveillance Information System, which records all cases of hospitalization due to Severe Acute Respiratory Syndrome (SARS), including COVID-19 infections, in Brazil. The algorithms were applied with the objective of identifying patterns and predicting the probability of mortality among hospitalized COVID-19 patients. In addition to predicting mortality, the study also sought to identify the most relevant variables for the outcome of death from COVID-19. For this, the calculation of variable importance was used through methods intrinsic to each algorithm and the SHAP technique (SHapley Additive Explanations), facilitating the interpretation of the models. The tested algorithms showed good levels of performance, achieving an accuracy of about 74% and an AUC-ROC of 74%. The results confirmed that elderly patients, patients who used ventilatory support, or those who required ICU admission were significantly more likely to die. These findings are consistent with the literature and reinforce the need for early and targeted interventions for these at-risk groups.

Keywords: COVID-19, Artificial Intelligence, Machine Learning, Risk Factors, Logistic Regression, Random Forest, Neural Networks.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

A pandemia de COVID-19 é um evento histórico sem precedentes que afetou a saúde e a vida de milhões de pessoas em todo o mundo. Desde o seu surgimento em dezembro de 2019 (HUANG C.; WANG, 2020), o novo coronavírus se espalhou rapidamente pelo globo, levando a medidas de contenção sem precedentes, como o distanciamento social, o fechamento de fronteiras e a interrupção de atividades econômicas. O impacto da COVID-19 foi sentido em todas as esferas da sociedade, desde a saúde pública até a economia, a educação e as relações sociais.

A pandemia de COVID-19 chegou ao Brasil em fevereiro de 2020, quando o país registrou seu primeiro caso confirmado da doença. Desde então, a COVID-19 se espalhou rapidamente por todo o território nacional, levando o país a se tornar um dos epicentros mundiais da pandemia.

Em março de 2020, o governo federal decretou estado de emergência em saúde pública de importância nacional em razão da pandemia de COVID-19. Desde então, várias medidas foram tomadas em todo o país para conter a disseminação do vírus, como a imposição de medidas de distanciamento social, o fechamento de escolas e comércios não essenciais, a restrição de viagens e a proibição de aglomerações.

No entanto, a implementação dessas medidas foi desigual em todo o país, com alguns estados e municípios adotando estratégias mais rigorosas e outros sendo mais lenientes. Além disso, o país enfrentou uma série de desafios na gestão da pandemia, como a escassez de equipamentos de proteção individual para profissionais de saúde, a falta de testes em massa e a falta de coordenação entre os governos federal, estaduais e municipais.

Neste trabalho iremos utilizar dados públicos, obtidos no Open Data SUS, sobre Síndrome Respiratória Aguda Grave e COVID para criar modelos de classificação que nos permitam estimar quais pacientes tem mais chance de vir a óbito com base em diversas características coletadas.

O objetivo é ter um entendimento da importância das variáveis utilizadas por modelos

de classificação a fim de obter maior explicabilidade dos modelos de inteligência artificial que têm ganhado muita notoriedade nos últimos anos com intuito de torná-los mais acessíveis ao público, ajudando assim na sua disseminação.

Para isso, consideramos apenas os casos de pacientes adultos, isto é, com 18 anos ou mais, hospitalizados por COVID-19 e que foram notificados no ano de 2022.

O questionário que monitora a Síndrome Respiratória Aguda Grave já existia antes da COVID, mas foi alterado em virtude desta. Essa mudança nos obriga a tomar um cuidado adicional na limpeza dos dados.

Os modelos serão criados usandos técnicas de aprendizado supervisionado com o objetivo de predizer o desfecho de um paciente internado com COVID-19. Com o objetivo de avaliar a capacidade preditiva dos modelos usaremos métricas de avaliações que envolvam sensibilidade e especificidade, como a acurácia, f1 score e área sob a curva ROC, entre outras que forem julgadas pertinentes.

Além do modelo preditivo em si, haverá um ganho de informação se conseguirmos extrair do modelo as informações referentes à importância das variáveis, a fim de identificar os fatores que mais influenciam no desfecho dos pacientes. Trabalhos semelhantes estão se tornando mais comuns como (DABBAGH, 2023) e (SILVADARCY RISOMARIO; NETO, 2022), e um entendimento dos modelos aumentará a usabilidade dos mesmos nas mais diversas áreas, trazendo benefícios e aumentando a responsabilidade de seus usos.

Em (SILVADARCY RISOMARIO; NETO, 2022), os autores usaram dados de pacientes diagnosticados com COVID-19 no Brasil no período de janeiro a setembro de 2021 e os algoritmos regressão logística, árvore de decisão e floresta aleatória para fazer a classificação de paciente mais propensos a vir a óbito. Não foi detalhado o tratamento e a limpeza de dados feita, mas foram utilizadas praticamente as mesmas variáveis, embora os autores não tenham considerado a região do Brasil e a escolaridade. No artigo também foi feito o estudo da importância das variáveis da floresta aleatória, que foi o melhor modelo preditivo obtido, e encontrou-se que as três principais variáveis são suporte ventilatório, UTI e idade.

Convém ressaltar que atualmente muitos estudos estão sendo realizados para entender como inteligência artificial e aprendizado de máquina podem se empregados na área da saúde, e muito desse esforço foi intensificado durante a pandemia da COVID-19. Apenas para citar alguns, (ZAERI, 2024) examinou um grande número de estudos e identificou áreas em que a inteligência artificial poderia ser empregada como, por exemplo, diagnóstico de pacientes atráves da análise de imagens de raio-X dos pulmões, análise de severidade e descoberta de possíveis novos medicamentos; (BAGABIR, 2022) utilizou inteligência artificial no estudo do genoma com fins de encontrar possíveis novos medicamentos e vacinas bem como novas variantes do coronavirus; (GUDIGAR, 2021) e (HUANG, 2021) usaram inteligência artificial para avaliar imagens de raio-X, tumografias computadorizadas e ultrassom para ajudar no

diagnóstico da doença; (COMITO CARMELA; PIZZUTI, 2022) conduziu uma revisão da literatura a fim de comparar diferentes métodos para previsão de pessoas infectadas com objetivo de auxiliar os sistemas de saúde a se preparar para a demanda adequada bem como técnicas de diagnóstico.

Muitos outros trabalhos poderiam ser citados e a existência de revistas especificamente voltadas para o tema de inteligência artificial na saúde como, por exemplo, a *Artificial Intelligence in Medicine* dá uma dimensão do crescimento e importância dessas técnicas.

É neste contexto que o presente trabalho se insere, analisando os impactos para o Brasil e focando na explicabilidade dos modelos. Somente com alguma interpretabilidade de modelos considerados "caixa-preta" é que estes poderão ser largamente adotados.

O objetivo deste trabalho não é o de apresentar um sistema que possa fazer triagem de pacientes de forma automática e substituir profissionais da saúde; pelo contrário, o objetivo é fornecer aos profissionais de saúde mais uma ferramenta de auxílio para realizar a triagem.

Capítulo 2

Referencial Teórico

Neste capítulo, apresentamos o histórico e o funcionamento dos modelos de classificação utilizados neste trabalho, a saber, regressão logística, floresta aleatória e redes neurais. Após a apresentação dos modelos discutimos também os valores SHAP, uma técnica que pode ser aplicada a qualquer modelo e nos dá uma medida do impacto de cada preditor na classificação.

Existe inúmeros modelos de classificação, mas usaremos apenas três neste trabalho: regressão logística, floresta aleatória e rede neural. Estes três foram escolhido por terem características bem distintas entre si e por serem modelos largamente utilizados em problema de classificação. A regressão logística é um modelo clássico, interpretável e não é computacionalmente intensivo; a floresta aleatória é conhecida por ter bons resultados usando dados tabulares; e a rede neural tem ganhado muita notoriedade na última década e se caracteriza por não ser explicável e ser mais computacionalmente intensiva.

Antes, porém, começaremos com uma descrição do conjunto de dados e dos tratamentos aplicados. Neste trabalho usou-se a linguagem Python (versão 3.11.3) e as bibliotecas seaborn (versão 0.12.2), scipy (versão 1.10.1), numpy (versão 1.23.5), matplotlib (versão 3.7.1), pandas (versão 1.5.3), shap (versão 0.46.0) e sklearn (versão 1.2.2), para analisar dados públicos sobre Síndrome Respiratória Aguda Grave disponibilizado no site do Open Data SUS.

A base de dados utilizada contém 556445 observações e 173 variáveis, conforme dicionário de dados disponibilizado no site do Open Data SUS. As variáveis utilizadas neste trabalho serão apresentadas na Tabela 3. Tais variáveis contém apenas informações sobre o paciente, tais como comorbidades e sintomas observados. A variável EVOLUÇÃO é a variável de interesse, pois ela contém informação sobre óbito ou cura do paciente.

2.1 Critérios de Seleção da Amostra

Primeiramente, precisamos fazer um trabalho de limpeza e pré processamento dos dados, além de uma análise exploratória. Para a variável EVOLUCAO descartou-se as observações que contenham "9 - Ignorado" pois não queremos imputar a variável alvo. Além disso, tratou-se as observações "2-Óbito" e "3-Óbito por outras causas" como "Óbito" Desse modo, o desfecho do estudo é uma variável binária, indicando a cura ou óbito do paciente. Iremos trabalhar apenas com os pacientes que foram notificados em 2022, essa escolha é feita para tentarmos pegar uma amostra mais homogênea, pois diferentes fases da pandemia tiveram interferência de diferentes fatores como, por exemplo, vacinação e melhor entendimento dos efeitos do vírus no organismo humano, e também para podermos comparar com trabalhos sobre o assunto que usaram bases de 2021.

Decidiu-se por analisar apenas pacientes internados, portanto a variável HOSPITAL inicialmente selecionada será desconsiderada, pois todos os pacientes analisados terão o valor 1 para essa variável tornando irrelevante em análises posteriores.

Iremos trabalhar somente com adultos, por isso selecionaremos a faixa de idade acima de 18 anos e iremos desconsiderar a variável TP_IDADE pois esta indica apenas se a idade é medida em anos, dias ou meses. Decidiu-se também trabalhar somente com pacientes que foram diagnosticados com COVID-19, portanto descartaremos a variável CLASSI_FIN após a seleção adequada.

2.2 Tratamento dos Dados

A base de dados possui muitos valores faltantes e muitas observações marcadas com "9 - Ignorado" para fatores associados, portanto decidiu-se considerar que tais observações indicam ausência do fator observado, pois devido ao grande número de atendimentos de triagem durante a pandemia, muitas equipes não possuíam tempo para preencher toda a ficha, geralmente, indicando apenas o que era observado no paciente. Dessa forma, de acordo com profissionais da área, substituiu-se dados faltantes e as observações não registradas para todas as variáveis associadas a comorbidades (puérpera, doença cardiovascular crônica, doença hematológica crônica, síndrome de Down, doença hepática crônica, asma, diabetes mellitus, doença neurológica crônica, outra pneumopatia crônica, imodeficiência ou imunosupressão, doença renal crônica, obesidade ou outros fatores) pela categoria 0.

Mesmo com tais escolhas para preencher dados faltantes, não temos como imputar valores para as variáveis CS_ESCOL_N (escolaridade), CS_SEXO, CS_RACA (raça), CS_ZONA (zona urbana/rural) e VACINA_COV (se a pessoa se vacinou para a COVID-19) então vamos trocar observações marcadas como ignorado para valores nulos. Essas observações serão posteriormente retiradas da análise para diminuir a incerteza sobre os dados disponíveis,

melhorando os resultados.

A variável SUPORT_VEN possuia originalmente três categorias ("1-Sim, invasivo", "2-Sim, não invasivo", "3-Não"). Neste caso, agrupamos os níveis "1-Sim, invasivo" e "2-Sim, não invasivo" em uma única categoria.

Também mudou-se a variável SG_UF_INTE que indica unidade federativa da internação do paciente para trabalharmos apenas com as cinco regiões do país, de acordo com a classificação do IBGE.

Após esses tratamentos iniciais verificou-se que na variável CS_ZONA não há observações para Rural, pois provavelmente essas observações foram descartadas em algum tratamento anterior, portanto vamos descartar toda essa variável.

Para a variável CS_RACA vamos considerar pretos e pardos como uma única categoria, pois mesmo sabendo que tais grupos possuem características únicas, o número de pardos é significativamente menor e em estudos internacionais tal consideração também costuma ser feita. Para a variável VACINA_COV vamos considerar que valores ignorados indicam falta de vacinação, isso também deve-se ao fato de o campo de marca da vacina estar nulo nestes casos, reforçando que o paciente não se vacinou.

Todas as mudanças e tratamentos descritos tiveram como objetivos limpar e padronizar os dados de acordo com um racional ao mesmo tempo em que pretendem reduzir ruídos e incertezas sobre os valores preenchidos nos formulários.

Por fim, vamos descartar todas as observações que ainda possuem dados faltantes. Isso nos deixa com 57016 observações e 36 variáveis, sendo um número suficiente para objetivos de encontrar modelos de classificação em momentos posteriores deste trabalho.

Após a limpeza ficamos com as variáveis: CS_SEXO, NU_IDADE_N, CS_RACA, CS_ESCOL_N, NOSOCOMIAL, FEBRE, TOSSE, GARGANTA, DISPNEIA, DESC_RESP, SATURACAO, DIARREIA, VOMITO, OUTRO_SIN, PUERPERA, CARDIOPATI, HEMATOLOGI, SIND_DOWN, HEPATICA, ASMA, DIABETES, NEUROLOGIC, PNEUMOPATI, IMUNODEPRE, RENAL, OBESIDADE, OUT_MORBI, UTI, SUPORT_VEN, EVOLUCAO, DOR_ABD, FADIGA, PERD_OLFT, PERD_PALA, VACINA_COV, REGIAO.

2.3 Regressão Logística

A regressão logística foi desenvolvida como um modelo de crescimento populacional em uma série de artigos do matemática belga Pierre François Verhulst.

Primeiramente Verhulst introduziu o conceito de função logística (VERHULST, 1838) e posteriormente mostrou como ajustar a curva à um conjunto de pontos (VERHULST, 1845).

A função logística, também chamada de sigmóide por seu gráfico lembrar um "S", é

definida por $f : \mathbb{R} \to (0, 1)$ em que $f(t) = \dfrac{1}{1 + e^{-t}}$ e foi redescoberta posteriormente e de forma independente por outras pessoas.

A regressão logística foi usada originalmente em problemas de classificação binária, isto é, quanto temos uma variável que assume apenas dois valores, habitualmente chamados de 0 (zero) representando fracasso e 1 (um) representando sucesso; embora possa ser adaptada para classificação com mais de dois valores e para outros cenários. Embora tenha originalmente aparecido nos trabalhos de Verhulst a regressão logística se popularizou como modelo estatístico com Joseph Berkson (BERKSON, 1944).

Na função logística, se t é combinação linear de alguma outra variável independente x, então $t = \beta_0 + \beta_1 x$, logo $p(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$. Neste caso, $p(x)$ é interpretado como a probabilidade da variável dependente Y ser igual a 1. Para o caso de múltiplas variáveis independentes temos $t = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$ e $p(x) = \dfrac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{m} \beta_i x_i)}}$

A estimação dos coeficientes β_i é feita via Máxima Verossimilhança e a vantagem deste modelo é que é possível realizar testes de hipóteses para verificar a significância dos valores estimados, bem como obter intervalos de confiança e interpretabilidade dos coeficientes, embora a interpretação dos coeficientes não seja tão direta como na regressão linear.

Uma das vantagens da regressão logística é permitir a interpretabilidade do modelo. Para interpretar os coeficientes de regressão, utilizamos a Razão de Chances (RC), $RC = \dfrac{\text{chance de ter efeito}}{\text{chance de não ter efeito}}$, isto é, ela calcula quantas vezes a chance de ter um efeito é maior do que a chance de não ter esse efeito. Lambrando que a chance é calculada por $\text{chance} = \dfrac{\text{probabilidade de ocorrência}}{\text{probabilidade de não ocorrência}}$. Valores de chance próximos de 1 indicam que as probabilidades de ter efeito e de não ter efeito são praticamente as mesmas.

Portanto, podemos entender a importância de cada variável na regressão logística via razão de chances. Na regressão logística e^{β_i} representa a mudança na razão de chances para uma mudança unitária no x_i. Por exemplo, se $e^{\beta_i} = 0,5$ e $x_i = 2$, então as chances de ocorrer sucesso é $0,5$ vezes a chance de sucesso para $x_i = 1$.

2.4 Floresta Aleatória

Proposto em 1995 (HO, 1995) esse modelo apresenta o nome de floresta pois é construído com base em vários modelos chamados árvores de decisão. Vamos ter um primeiro entendimento de como funcionam as árvores de decisão e então passaremos para a floresta aleatória.

2.4.1 Árvore de Decisão

Uma árvore de decisão é um modelo construído com base em condicionais, isto é, se algo acontecer decide-se por uma ação, caso contrário decide-se por outra. Por exemplo, se uma pessoa quer ir à praia apenas em dias ensolarados e sem vento e encontra-se em um momento de tomada de decisão sobre ir ou não à praia. Neste caso, se não estiver ensolarado a pessoa não irá à praia; se estiver ensolarado então deve-se observar se há vento, pois não havendo vento irá à praia e havendo vento não irá à praia.

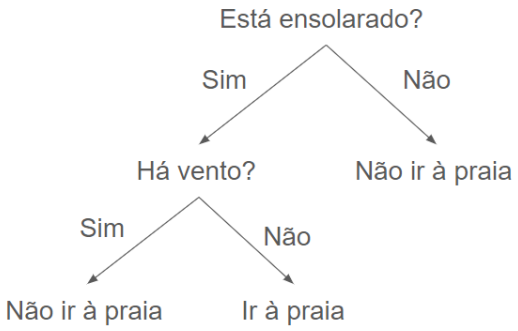

Figura 1 – Exemplo de Árvore de Decisão

Costumamos tomar decisões deste tipo todos dias, e a forma da tomada de decisão pode ser apresentada de forma gráfica, lembrando uma árvore. Em geral, desenha-se um nó a partir do qual partem dois caminhos e cada caminho se ramifica novamente e assim por diante.

Qual é a diferença do nosso processo de tomada de decisão para uma árvore de decisão? Quando tomamos decisões nós costumamos estabelecer os critérios em que haverá bifurcações com relação à tomada de decisão, e quando usamos uma árvore de decisão os critérios são escolhidos pelo modelo de forma a conseguir fazer a melhor separação de classe possível com base em algum critério como Índice de Gini ou entropia.

A divisão dos ramos de uma árvore de decisão pode ser feita utilizando diferentes critérios de impureza, sendo os mais comuns o Índice de Gini e a Entropia. Ambos avaliam o quão "puro"ou "impuro"é um conjunto de dados, ou seja, quão misturadas estão as classes nos ramos resultantes da divisão.

Vamos entender melhor como é feita a divisão dos nós numa árvore de decisão, vamos começar com o índice de Gini.

2.4.2 Índice de Gini

O Índice de Gini mede a probabilidade de classificar incorretamente um item escolhido aleatoriamente se ele for rotulado de acordo com a distribuição de classes no nó. Quanto menor o valor do Gini, mais "puro"é o nó, isto é, as classes estão mais homogêneas e, portanto, melhor foi a separação das classe.

A fórmula do Índice de Gini é:

$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2$ em que p_i é a proporção de instâncias da classe i no conjunto e n é o número total de classes.

O critério de divisão usando o Índice de Gini busca minimizar a impureza resultante, ou seja, encontrar o ponto de divisão que resulta nos nós mais homogêneos.

2.4.3 Entropia

A Entropia é outra medida de impureza, originada da teoria da informação. Ela quantifica a incerteza associada à previsão da classe de um elemento no conjunto. Se todos os elementos pertencem à mesma classe, a entropia é 0 (mínima); se as classes são distribuídas igualmente, a entropia é máxima.

A entropia é definida por $Entropia(p) = -\sum_{i=1}^{n} p_i \log_2(p_i)$, em que p_i é a proporção de instâncias da classe i no conjunto e n é o número de classes.

Dessa forma, o Índice de Gini ou a entropia são calculados para todas as variáveis a fim de escolher a que consegue fazer a melhor separação de classe. A variável escolhida é usada no nó, entao cada ramificação da árvore da origem a dois nós e as variáveis usadas em cada um é feita novamente observando-se quais apresentam a melhor separação de classes e assim sucessivamente.

2.4.4 Floresta Aleatória

Agora que entendemos intuitivamente como funciona uma árvore de decisão vamos partir para a construção da floresta. Como o nome sugere iremos usar várias árvores, mas as árvores não podem ser iguais, caso contrário todas as árvores iriam ter o mesmo comportamento e o conceito de floresta deixaria de fazer sentido. Para criar árvores que sejam preditores independentes iremos apresentar para cada árvore apenas um subconjuntos das variáveis, escolhidas de forma aleatória, e apenas um subconjuntos das observações. Assim, mesmo que duas árvores usem as mesmas variáveis, não terão exatamente o mesmo comportamento por serem construídas baseadas em observações diferentes.

A classificação feita por uma árvore é um procedimento relativamente simples. Basicamente a árvore usa as condições que definem os nós para tomar decisões do tipo "se condição, então ação" de modo a partir do nós central e ir descendo a árvore até as folhas. Quando se chega em uma folha a árvore classifica de acordo com a classe que compõe a maioria naquela folha.

Agora que entendemos como ela é construída precisamos entender como a decisão é tomada. Em problemas de classificação pode-se atribuir classificação à uma variável tomando a classificação mais frequente para cada árvore, numa espécie de votação, e em problemas de

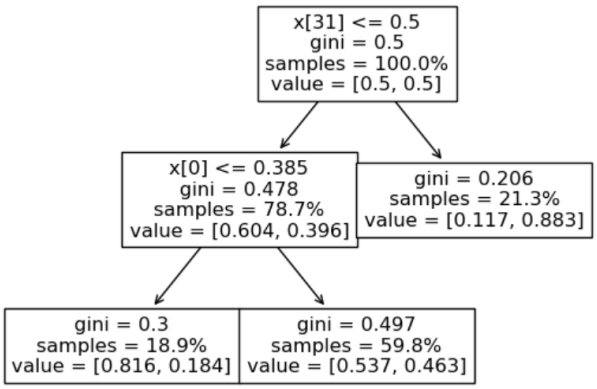

Figura 2 – Exemplo de Árvore de Decisão

regressão pode-se considerar as médias dos valores de cada árvore.

A importância das variáveis na floresta aleatória é feita considerando-se a redução do Gini ou da entropia, medindo a contribuição de cada variável para a homogeneidade dos nós e folhas na árvore de decisão.

Cada vez que uma variável X_i é usada para dividir um nó em uma árvore, a impureza do nó é reduzida. A importância do nó é então calculada somando todas as reduções do Gini ou da entropia proporcionada pela variável X_i em todas as árvores da floresta. Essa é uma maneira intrínseca de medir a importância de variáveis em algoritmos baseados em árvore.

2.5 Rede Neural

Criada para imitar o funcionamento do cérebro humano por Warren McCulloch e Walter Pitts em 1943 (MCCULLOCH, 1943) a rede neural considera vários neurônios artificiais agrupados em camada que são interligadas. A Figura 3 ilustra uma rede neural do tipo *Multi Layer Perceptron* (MLP), semelhante à que será usada neste trabalho.

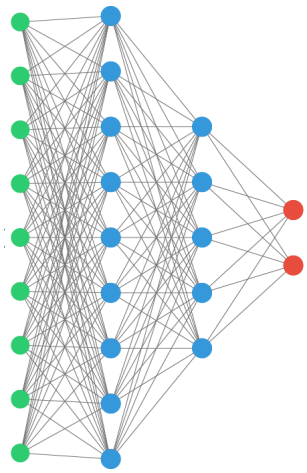

Figura 3 – Rede Neural do tipo Multi Layer Perceptron

Na Figura 3, temos a camada de entrada à esquerda e a camada de saída à direita, todas as demais são chamadas de camadas ocultas. O termo Aprendizagem Profunda (*Deep Learning*, em inglês) refere-se à redes neurais com muitas camadas ocultas, em geral mais do que três, embora os modelos que deram popularidade às redes neurais e estão muito conhecidos atualmente, como o ChatGPT (BASTIAN, 2023), tenham dezenas de camadas, cada uma com centenas de neurônios, chegando a mais de 175 bilhões de parâmetros.

Cada neurônio da camada de entrada vai assumir um valor de uma variável, os neurônios seguintes recebem combinações lineares dos valores de entrada vezes os pesos de cada neurônio da camada anterior e aplica-se a esse valor uma função conhecida como função de ativação, a fim de introduzir não linearidade na rede. Também há uma constante chamada viés que é somada à combinação linear.

Existem muitas funções de ativação disponíveis, neste trabalho usaremos a função ReLu. O nome ReLu é um acrônimo de *Rectified Linear Unit*, e é muito usada por ser computacionalmente mais eficiente. É definida por $f(x) = \max(x, 0)$.

Os pesos w e o viés b são parâmetros ajustáveis que determinam a saída de uma rede neural, e o objetivo é ajustá-los de forma que a previsão da rede $\hat{y}$ seja a mais próxima possível do valor desejado y. Esses parâmetros são inicializados de forma aleatória no início do treinamento, e o processo de aprendizado consiste em ajustá-los para minimizar o erro de predição.

A rede neural é composta por camadas de neurônios, onde cada neurônio em uma camada é conectado aos neurônios da camada seguinte. A equação básica que descreve o comportamento de um neurônio pode ser expressa como $z = \sum_{i=1}^{n} w_i x_i + b$, em que x_i são as variáveis, w_i são os pesos associados a cada variável x_i, e o viés b é um valor que ajusta a saída de um neurônio independentemente das entradas.

O valor z resultante dessa soma ponderada das entradas é então passado por uma função de ativação $f(z)$, que adiciona não-linearidade ao modelo.

A saída do neurônio $f(z)$ é então passada para a próxima camada da rede, e esse processo é repetido até a camada de saída, onde a previsão final $\hat{y}$ é feita.

Para que a rede neural aprenda, ela precisa ajustar seus pesos e viés de forma a minimizar a diferença entre a previsão $\hat{y}$ e o valor esperado y. Essa diferença é medida por uma função de custo (ou função de perda), entropia cruzada para problemas de classificação: $J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$, em que m é o número de exemplos no conjunto de dados, $\hat{y}_i$ é a previsão feita pela rede para o exemplo i, e y_i é o valor real esperado para o exemplo i.

Para minimizar a função de custo, utilizamos o gradiente descendente. O gradiente descendente é uma técnica iterativa que ajusta os pesos e o viés da rede na direção oposta ao gradiente da função de custo em relação a esses parâmetros. O gradiente $\partial J(\theta)$ é o vetor de

derivadas parciais da função de custo em relação a cada peso e viés da rede

$$\partial J(\theta) = \left(\frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \ldots, \frac{\partial J}{\partial w_m}, \frac{\partial J}{\partial b} \right)$$

.

O ajuste dos pesos e viés é feito de acordo com a seguinte regra de atualização $w_i \leftarrow w_i - \eta \frac{\partial J}{\partial w_i}$ e $b \leftarrow b - \eta \frac{\partial J}{\partial b}$ em que η é a taxa de aprendizado, um hiperparâmetro que controla o tamanho do passo na direção do gradiente, e $\frac{\partial J}{\partial w_i}$ e $\frac{\partial J}{\partial b}$ são as derivadas parciais da função de custo em relação ao peso e viés, respectivamente. Este processo é repetido para cada amostra do conjunto de dados durante o treinamento. Uma época é definida como um ciclo completo em que todos os exemplos de treinamento são usados para ajustar os pesos e vieses.

Backpropagation (retropropagação) é o algoritmo utilizado para calcular os gradientes da função de custo em relação a todos os pesos da rede neural. Ela funciona propagando o erro da previsão final $\hat{y}$ de volta para as camadas anteriores da rede, utilizando a regra da cadeia para calcular as derivadas parciais. A retropropagação atualiza os pesos de cada camada de modo a minimizar o erro da previsão para o próximo passo, da seguinte forma: $\frac{\partial J}{\partial w_i} = \frac{\partial J}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial w_i}$

Esse cálculo é feito em todas as camadas da rede, permitindo que os pesos em todas as camadas sejam ajustados.

O processo de ajuste de pesos e vieses via gradiente descendente continua até que a função de custo atinja um mínimo. Esse mínimo pode ser um mínimo local (para redes mais complexas) ou o mínimo global, dependendo da superfície da função de custo e da estrutura do modelo. A qualidade do ajuste depende, entre outros fatores, da taxa de aprendizado e do número de épocas.

Usualmente, uma rede é treinada com muitas épocas e há ainda parâmetros que poderiam ser acrescentados e/ou alterados para criar uma rede neural.

2.6 Validação Cruzada

A validação cruzada é uma técnica de avaliação de modelos que divide um conjunto de dados aleatoriamente em n partes ou "*folds*"iguais.

A seguir, o modelo é treinado e avaliado várias vezes, cada vez usando um *fold* diferente como conjunto de teste e os demais *folds* como conjunto de treinamento. Isso resulta em n iterações:

Na primeira iteração, o primeiro *fold* é usado para teste, e os demais são usados para treinamento. Na segunda iteração, o segundo *fold* é usado para teste, e os demais são usados para treinamento. E assim sucessivamente, até que cada *fold* tenha sido usado uma vez como conjunto de teste.

Para cada iteração, uma métrica de desempenho como, por exemplo, acurácia é calculada com base no conjunto de teste. Após as n iterações, as métricas de desempenho são calculadas como a média das n iterações, fornecendo uma estimativa mais robusta do desempenho do modelo, pois não depende de uma amostra específica.

A validação cruzada com n folds é especialmente útil quando se tem um conjunto de dados limitado, pois garante que cada ponto de dado seja usado tanto para treinamento quanto para teste, ajudando a avaliar o modelo de maneira mais consistente e a evitar overfitting.

2.7 Métricas de Qualidade Preditiva

Neste trabalho usou-se diferentes métricas para medir a qualidade preditiva dos modelos de classificação: acurácia, *recall* (também conhecida como sensibilidade), *precision*, *f1-score* e ROC-AUC. Vamos explicar como são calculadas essas métricas.

Ao realizar classificação binária há quatro situações possíveis de predição:

1. modelo prevê 1 e o resultado verdadeiro é 1 (verdadeiro positivo)
2. modelo prevê 1 e o resultado verdadeiro é 0 (falso positivo)
3. modelo prevê 0 e o resultado verdadeiro é 1 (falso negativo)
4. modelo prevê 0 e o resultado verdadeiro é 0 (verdadeiro negativo)

Temos assim, duas possibilidade de acerto e duas de erro para cada previsão. Para facilitar esse entendimento, essas quatro possibilidade são exibidas em uma matriz, conhecida como matriz de confusão.

Tabela 1 – Matriz de Confusão

Real	Previsto	
	Sim	Não
Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Dessa forma, quando um modelo faz previsões para várias observações, quanto maiores forem o verdadeiro positivo e o verdadeiro negativo e quanto menor forem o falso positivo e o falso negativo, melhor está o modelo. No entanto, em muitos casos não é possível obter altos valores de verdadeiro positivo e verdadeiro negativo de tal forma que devemos priorizar um deles.

Para considerar todos esses casos usamos mais de uma métrica, para uma melhor compreensão da qualidade do modelo como um todo. Dessa forma, partindo das quatro possibilidade que temos para cada previsão, temos:

- Acurácia $= \frac{VP + VN}{VP + VN + FP + FN}$
- *Recall* $= \frac{VP}{VP + FN}$
- *Precision* $= \frac{VP}{VP + FP}$
- *f1-score* $= 2 * \frac{Precision * Recall}{Precision + Recall}$

Em palavras, a acurácia indica uma performance geral do modelo, isto é, o percentual de previsões feitas corretamente, dentre todas as previsões; *precision* indica o percentual de previsões corretas dentre todas as previsões da classe positiva; *recall* é o percentual de previsões corretas dentre todos os casos da classe positiva; e o *f1-score* é a média harmônica entre *precision* e *recall*.

Caso as classes sejam desbalanceadas, isto é, se houver muitos mais observações de uma classe do que da outra, e o modelo prever sempre a classe predominante, o resultado será uma alta acurácia, apesar do modelo ser ruim e não poder ser utilizado. Casos semelhantes podem acontecer com *precision* e *recall*, por isso é importante levar todas as métricas em consideração. O *f1-score*, por ser a média harmônica entre *precision* e *recall*, vai diminuir se uma das métrica estiver baixa, mesmo com a outra alta, assim é um indicativo de que ambas estão altas.

Outra métrica utilizada é ROC-AUC, em que ROC significa *receiver operating characteristic curve* e AUC significa *area under the curve*. Esse nome vem de uma curva que se obtém ao plotar a taxa de verdadeiros positivos contra a taxa de falsos positivos, que é conhecida como curva ROC. Ver Figura 12 para melhor entendimento. A área sob esta curva dá uma ideia da qualidade do modelo, pois quanto mais próximos de 1, melhor o modelo e quanto mais próximo de 0.5, mais o modelo se aproxima de um preditor aleatório.

2.8 Valores SHAP

SHAP (*SHapley Additive exPlanations*) é uma abordagem para explicar a importância das variáveis de qualquer modelo de aprendizado de máquina, introduzido em 2017 (LUNDBERG S. M.; LEE, 2017) e fundamenta-se em conceitos de teoria dos jogos que remontam ao artigo A Value for n-Person Games publicado em 1953 por Lloyd S. Shapley.

SHAP atribui a importância das variáveis de forma mais ponderada, considerando todas as possíveis combinações de variáveis. Em outras palavras, SHAP mede a contribuição marginal de cada variável no impacto total da "assertividade" do modelo. Este método é particularmente poderoso porque fornece explicações consistentes e interpretáveis.

Note que, apresentamos uma forma de calcular a importância das variáveis de um modelo de regressão logística e de um modelo de floresta aleatória, mas seria um erro dizer que um dos modelos dá mais importância para uma variável X_i somente baseados nos métodos intrínsecos de importância apresentados para cada modelo. Isso ocorre porque o método de calcular a importância é diferente, tornando a escala de variação diferente.

Além disso, nem todo modelo tem um método intrínseco de interpretação de importância de variáveis, como é o caso das redes neurais. Em vista disso, uma abordagem agnóstica aos modelos é interessante, e é por isso que usar valores SHAP é tão interessante, além de nos permitir ter uma interpretação natural.

Vamos entender como calcular o valor de Shapley. Para um conjunto de variáveis $X = \{x_1, x_2, \ldots, x_p\}$ e uma função de predição f, o valor de Shapley para uma variável x_i é dada por

$$\phi_i = \sum_{S \subset X \setminus \{x_i\}} \frac{|S|!(|X| - |S| - 1)!}{|X|!} [f(S \cap \{x_i\}) - f(S)]$$

em que:

- S é qualquer subconjunto das variáveis que não inclui x_i;
- $|S|$ é o número de variáveis no subconjunto S;
- $f(S)$ é a predição do modelo com as variáveis em S.

Os valores SHAP podem ser interpretados como a contribuição de cada variável para a predição de um dado exemplo. Um valor SHAP positivo para uma variável indica que essa variável contribui para obter uma classe positiva, enquanto um valor negativo indica que a variável contribui para obter a classe negativa.

Vejamos um exemplo para facilitar o entendimento do cálculo dos valores SHAP.

Suponha que três jogadores, A, B e C, estejam colaborando em um jogo para ganhar um prêmio, e suas contribuições não são iguais. Se eles vencerem a competição, como deveriam dividir o prêmio proporcionalmente à contribuição de cada um?

Vamos considerar que temos acesso à quantidade de pontos que os jogadores A, B e C obtêm quando jogam em todos os cenários possíveis, considerando a Tabela 2.

Vamos calcular a contribuição do jogador A. Considere que a função $\nu()$ calcula os pontos obtidos por um subconjunto de jogadores e vamos considerar todos os cenários possíveis.

1. todos jogam sozinhos, neste caso $\nu(A) = 10$;

Tabela 2 – Tabela com Pontuações para Exemplificar o Cálculo de Valores SHAP

Jogadores	Pontos
A	10
B	30
C	40
A, B	35
A, C	45
B, C	50
A, B, C	100

2. A joga sozinho, mas B e C jogam juntos, nesse caso $\nu(A) = 10$;
3. A joga com B, nesse caso $\nu(A, B) - \nu(B) = 35 - 30 = 5$;
4. A joga com C, nesse caso $\nu(A, C) - \nu(c) = 45 - 40 = 5$;
5. A joga com B, então C se junta a eles, nesse caso $\nu(A, B, C) - \nu(B, C) = 100 - 50 = 50$;
6. A joga com C, então B se junta a eles, nesse caso $\nu(A, B, C) - \nu(B, C) = 100 - 50 = 50$.

Portanto, o valor de Shapley para A é dado por $\frac{10+10+5+5+50+50}{6} = 21,67$

Para avaliar a importância de variáveis em modelos de aprendizado de máquina nós devemos trocar os jogadores por variáveis e a pontuação do jogo por valores em uma função de interesse como, por exemplo, a acurácia.

Podemos medir a acurácia do modelo com todas as possíveis combinações de variáveis e assim os valores SHAP nos ajudarão a entender quais variáveis mais contribuem para obter um bom valor para a função de interesse e, portanto, temos uma ideia de importância de variáveis através das suas contribuições marginais.

Vale ressaltar que como são feitas todas as combinações possíveis de variáveis para medir a contribuição marginal de cada uma, esse método se torna computacionalmente intensivo se o número de variáveis for elevado.

Este não é um método de seleção de variáveis, embora possamos a posteriori selecionar apenas as variáveis com valores SHAP relevantes.

Capítulo 3

Resultados

Tendo feito o tratamento inicial dos dados e definido os modelos que usaremos neste trabalho, vamos descrever os resultados encontrados.

3.1 Análise Exploratória

Inicialmente, realizamos uma análise descritiva da base de dados. Os resultados da frequência absoluta e relativa de cada variável em função da evolução do paciente, bem como o p-valor do teste qui-quadrado de Pearson são apresentados na Tabela 3, esse teste foi feito para verificar se existe associação significativa entre o preditor e a evolução.

Observando-se a coluna Total da Tabela 3, vemos que a amostra está balanceada em relaçao ao sexo mas desbalanceada em relação às demais variáveis, nota-se ainda um percentual pouco maior de óbito para pessoas do sexo masculino.

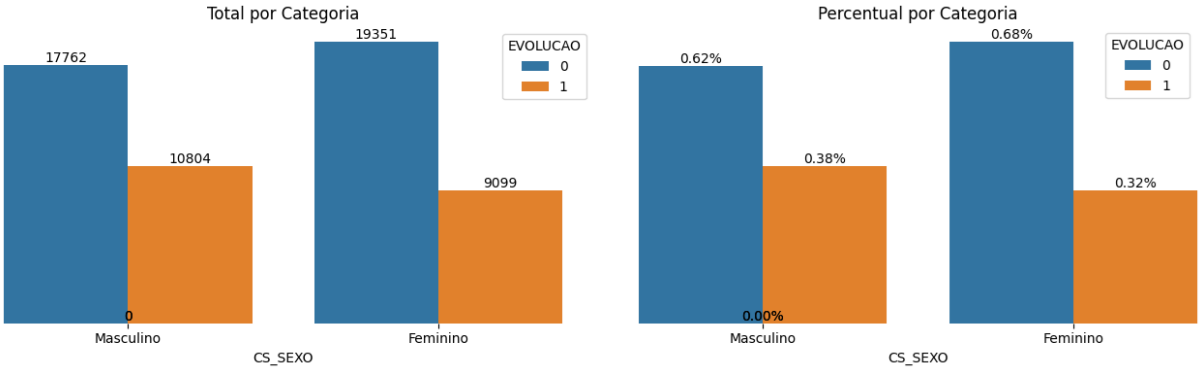

Figura 4 – Quantidade de Óbitos Separado por Sexo.

Apesar de haver uma grande diferença no valor absoluto de pacientes indicados com diferentes raças, pode-se notar que o percentual de óbito é praticamente o mesmo para todas as raças. It is interesting that the percentual of mortes is higher for pessoas with the highest escolaridad.

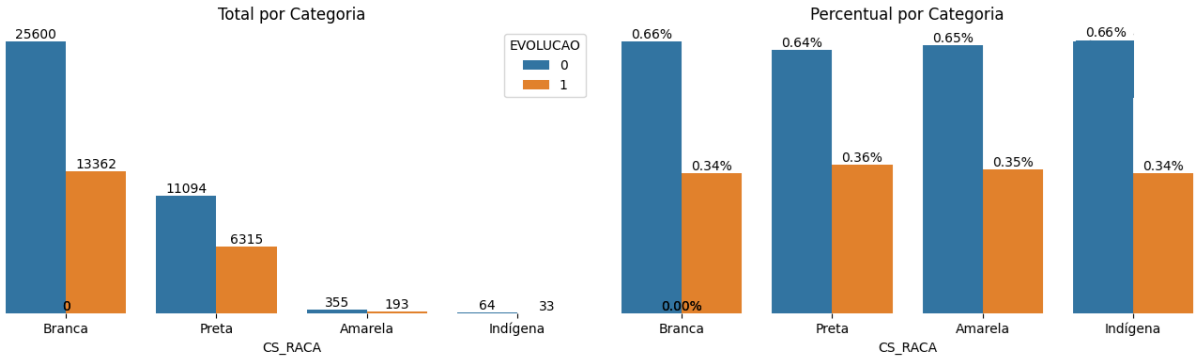

Figura 5 – Quantidade de Óbitos Separado por Raça.

Nota-se também um maior percentual de óbito dentre as pessoas que relataram dispnéia, em relação às que não relataram. O mesmo pode ser notado para pessoas que relataram desconforto respiratório ou que apresentaram concentração de oxigênio menor que 95%.

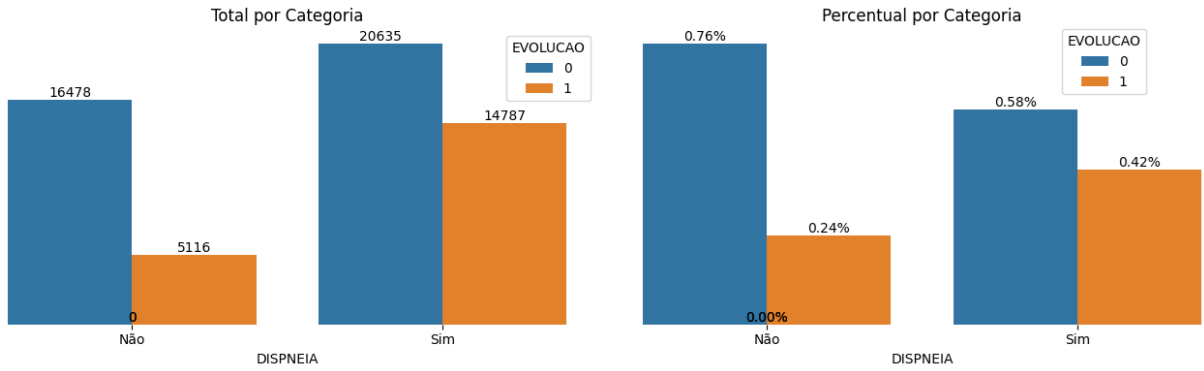

Figura 6 – Quantidade de Óbitos Separado por Dispneia.

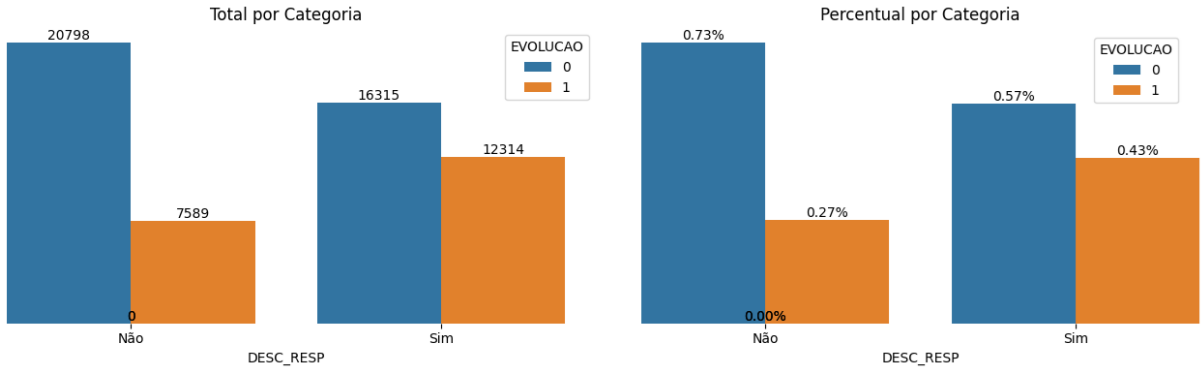

Figura 7 – Quantidade de Óbitos Separado por Desconforto Respiratório.

Podemos ver que há pouca puérperas com a doença e o percentual dessas que vieram a óbito é muito menor, isto provavalmente deve-se ao fato de que foram pessoas que tiveram um bom acompanhamento durante a doença.

Pessoas com doenças crônicas como, doença cardiovascular, doença hematológica, doença neurológica, doença renal, doença hepática ou outra pneumopatia, apresentaram percentuais maiores de óbito em relação às pessoas que não possuiam essas doenças.

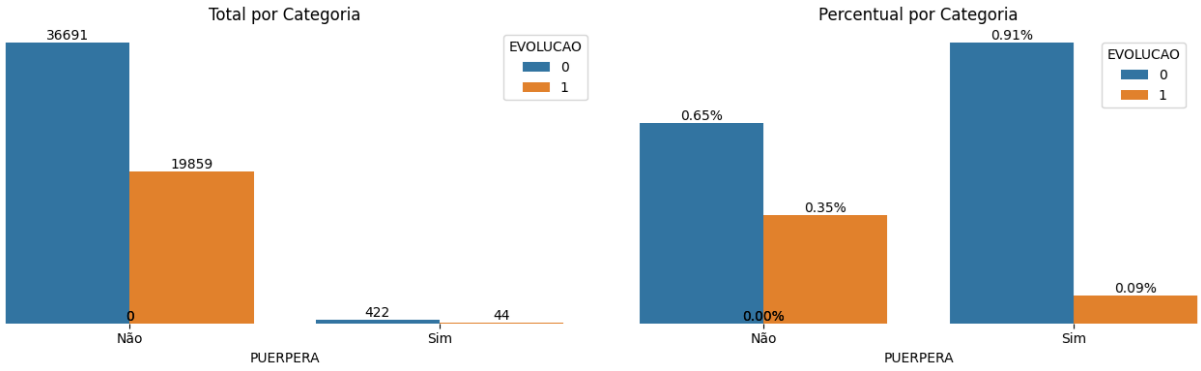

Figura 8 – Quantidade de Óbitos Separado por Puerpera.

Observa-se que as pessoas que receberam suporte ventilatório tiveram um maior percentual de óbitos, isso é razoável pois recebia o suporte pacientes que estavam em estado mais grave. Isso evidencia a importância de cuidados nas fases anteriores da doença para que esta não progredisse para estados mais graves.

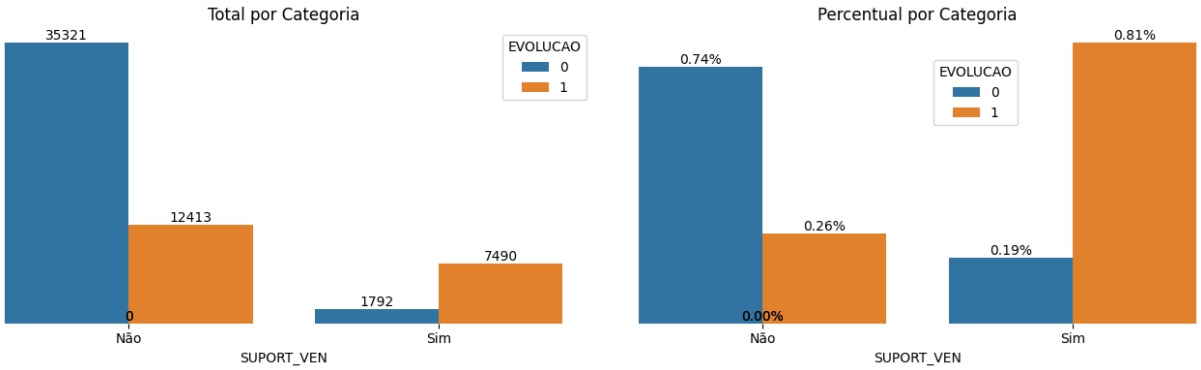

Figura 9 – Quantidade de Óbitos Separado por Suporte Ventilatório.

Além disso, o percentual de morte entre as pessoas que recebem a vacina da COVID-19 é menor se comparado ao grupo das pessoas não vacinadas.

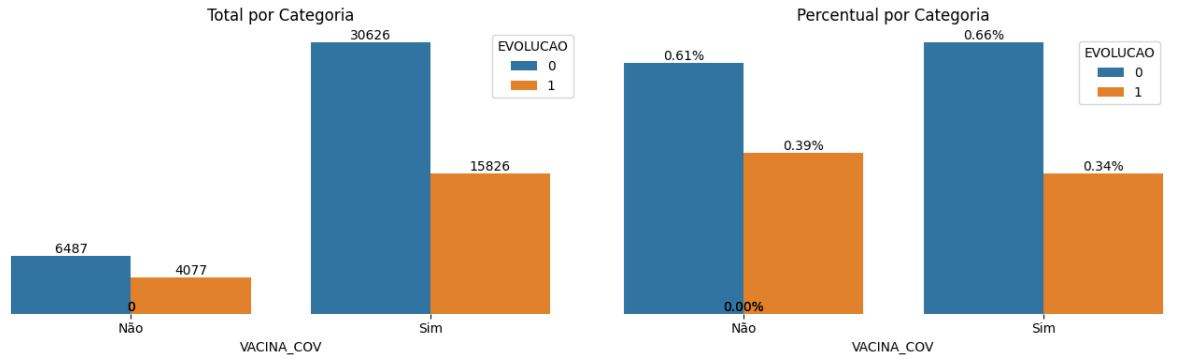

Figura 10 – Quantidade de Óbitos Separado por Vacina.

Conforme pode ser visto na Tabela 3, com exceção de Síndrome de Down, obesidade e fadiga, todos as demais variáveis apresentam-se relevantes ao nível de 5% de significância.

Tabela 3 – Frequências absolutas (relativas, %) e p-valor do teste qui-quadrado de Pearson.

Variável	Categoria	Cura	Óbito	Total	p valor
Sexo	Masc.	17762 (62)	10804 (38)	28566	<0,001
	Fem.	19351 (68)	9099 (32)	28450	
Raça	Branca	25600 (66)	13362 (34)	38962	<0,001
	Preta/Parda	11094 (64)	6315 (36)	17409	
	Amarela	355 (65)	193 (35)	548	
	Indígena	64 (66)	33 (34)	97	
Escolaridade	Analfabeto	10573 (62)	6559 (38)	17132	<0,001
	Fundamental	12237 (61)	7811 (39)	20048	
	Médio	9586 (71)	3902 (29)	13470	
	Superior	4717 (74)	1631 (26)	6348	
Nosocomial	Não	35410 (65)	18751 (35)	54161	<0,001
	Sim	1703 (60)	1152 (40)	2855	
Febre	Não	19921 (63)	11535 (37)	31456	<0,001
	Sim	17192 (67)	8368 (33)	25560	
Tosse	Não	11548 (59)	8112 (41)	19660	<0,001
	Sim	25565 (68)	11791 (32)	37356	
Dispneia	Não	16478 (76)	5116 (24)	21594	<0,001
	Sim	20635 (58)	14787 (42)	35422	
Desconforto Respiratório	Não	20798 (73)	7589 (27)	28387	<0,001
	Sim	16315 (57)	12314 (43)	28629	
Saturação $O_2 < 95\%$	Não	18202 (77)	5337 (23)	23539	<0,001
	Sim	18911 (56)	14566 (44)	33477	
Diarreia	Não	33568 (65)	18129 (35)	51697	0.013
	Sim	3545 (67)	1774 (33)	5319	
Vômito	Não	33640 (65)	18393 (35)	52033	<0,001
	Sim	3473 (70)	1510 (30)	4983	
Outros Sintomas	Não	24897 (64)	14257 (36)	39154	<0,001
	Sim	12216 (68)	5646 (32)	17862	
Puérpera	Não	36691 (65)	19859 (35)	56550	<0,001
	Sim	422 (91)	44 (9)	466	
Doença Cardiovascular Crônica	Não	23427 (68)	10802 (32)	34229	<0,001
	Sim	13686 (60)	9101 (40)	22787	
Doença Hematológica Crônica	Não	36636 (65)	19541 (35)	56177	<0,001
	Sim	477 (57)	362 (43)	839	

Variável	Categoria	Cura	Óbito	Total	p valor
Síndrome de Down	Não	36990 (65)	19831 (35)	56821	0.61
	Sim	123 (63)	72 (37)	195	
Doença Hepática Crônica	Não	36716 (65)	19421 (35)	56137	<0,001
	Sim	397 (45)	482 (55)	879	
Asma	Não	35938 (65)	19454 (35)	55392	<0,001
	Sim	1175 (72)	449 (28)	1624	
Diabetes mellitus	Não	28112 (67)	14072 (33)	42184	<0,001
	Sim	9001 (61)	5831 (39)	14832	
Doença Neurológica Crônica	Não	34488 (66)	17590 (34)	52078	<0,001
	Sim	2625 (53)	2313 (47)	4938	
Outra Pneumopatia Crônica	Não	37777 (66)	17966 (34)	55743	<0,001
	Sim	2336 (55)	1937 (45)	4273	
Imunodepressão	Não	35203 (66)	18200 (34)	53403	<0,001
	Sim	1910 (53)	1703 (47)	3613	
Doença Renal Crônica	Nâo	35111 (66)	17978 (34)	53089	<0,001
	Sim	2002 (51)	1925 (49)	3927	
Obesidade	Não	35058 (65)	18734 (35)	53792	0.10
	Sim	2055 (64)	1169 (36)	3224	
Outros Fatores de Risco	Não	24626 (68)	11429 (32)	36055	<0,001
	Sim	12487 (60)	8474 (40)	20961	
Região do Brasil	Sul	9245 (70)	3830 (30)	13075	<0,001
	Sudeste	18861 (63)	11225 (37)	30086	
	Centro Oeste	2886 (70)	1246 (30)	4132	
	Norte	2578 (66)	1309 (34)	3887	
	Nordeste	3543 (61)	2293 (39)	5836	
Suporte Ventilatório	Não	35321 (74)	12413 (26)	47734	<0,001
	Sim	1792 (19)	7490 (81)	9282	
Dor Abdominal	Não	34063 (65)	18621 (35)	52684	<0,001
	Sim	3050 (70)	1282 (30)	4332	
Fadiga	Não	27527 (65)	14860 (35)	42387	0.20
	Sim	9586 (66)	5043 (34)	14629	
Perda do Olfato	Não	35694 (65)	19409 (35)	55103	<0,001
	Sim	1419 (74)	494 (26)	1913	
Perda do Paladar	Não	35559 (65)	19360 (35)	54919	<0,001
	Sim	1554 (74)	543 (26)	2097	
Recebeu Vacina COVID-19	Não	6487 (61)	4077 (39)	10564	<0,001
	Sim	30626 (66)	15826 (34)	46452	

3.2 Regressão Logística

Para o treinamento da regressão logística usou-se a biblioteca statsmodels com a otimização de hiperparâmetros feita pela biblioteca optuna. Após algumas iterações com uso de validação cruzada, usando 5 *folds*, encontrou-se como melhor resultado.

Optamos por usar 5 *folds* em cada modelo pois temos um bom número de observações e esse valor não é tão computacionalmente intensivo e garante alguma robustez para melhor avaliação do modelo.

Na Figura 11 encontramos a matriz de confusão para a regressão logística, e na Figura 12 encontramos sua curva ROC.

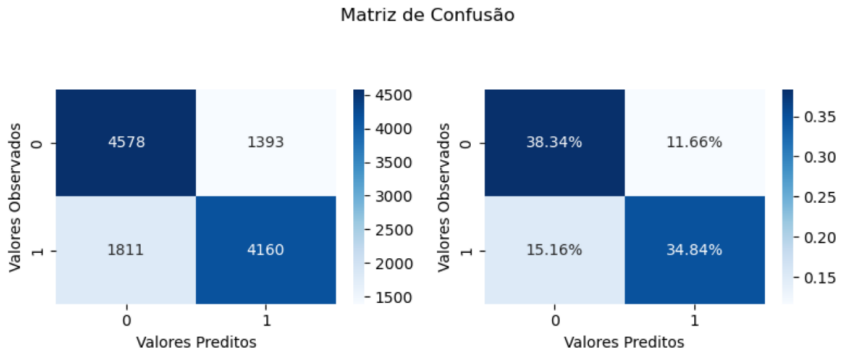

Figura 11 – Matriz de Confusão para a Regressão Logística

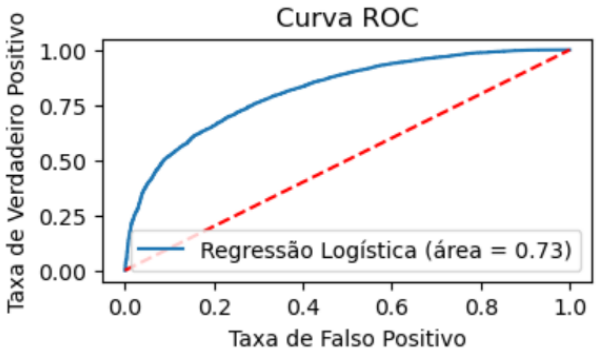

Figura 12 – Curva ROC para a Regressão Logística

Os principais valores de coeficientes da regressão logística, aqueles que possuem coeficientes com maiores valores absolutos, podem ser encontrados na Tabela 4.

Avaliando os valores da Tabela 4 notamos que o Suporte ventilatório é a variável que mais influencia a probabilidade de óbito, de modo que a chance de um paciente que usou suporte ventilatório vir a óbito é quase 8 vezes a chance de um paciente que não utilizou. Observamos, também que a chance de óbito é maior em pacientes que foram para a UTI, pacientes imunodepressivos e em pacientes com doenças hepática, renal, neurológica ou hematológica crônica. Por outro lado, pacientes que receveram vacina, tem asma, apresentaram tosse ou puérperas tem menor chance de óbito.

3.3 Floresta Aleatória

Para o treinamento da floresta aleatória usou-se a biblioteca scikit-learn, fizemos uma otimização de hiperparâmetros no espaço paramétrico em que variou-se o número de árvores, a

Tabela 4 – Principais Coeficientes da Regressão Logística

Preditor	Categoria	Estimativas	Chance	pvalor
Idade		3,36	28,79	$< 0,001$
Suporte Ventilatório	Sim	2,08	8,00	$< 0,001$
Imunodepressão	Sim	0,80	2,23	$< 0,001$
UTI	Sim	0,75	2,12	$< 0,001$
Doença Hepática Crônica	Sim	0,72	2,05	$< 0,001$
Doença Renal Crônica	Sim	0,46	1,58	$< 0,001$
Doença Neurológica Crônica	Sim	0,41	1,51	$< 0,001$
Doença Hematológica Crônica	Sim	0,36	1,43	$< 0,001$
Recebeu Vacina COVID-19	Sim	-0,33	0,72	$< 0,001$
Asma	Sim	-0,40	0,67	$< 0,001$
Tosse	Sim	-0,43	0,65	$< 0,001$
Puérpera	Sim	-0,66	0,52	$< 0,001$

profundidade das árvores e a quantidade de variáveis usada em cada árvores. Após 50 iterações com uso de validação cruzada, usando 5 *folds*, encontrou-se como melhor resultado usar 150 árvores, uma profundidade máxima de 10 nós.

Na Figura 13 encontramos a matriz de confusão para a regressão logística, e na Figura 14 encontramos sua curva ROC.

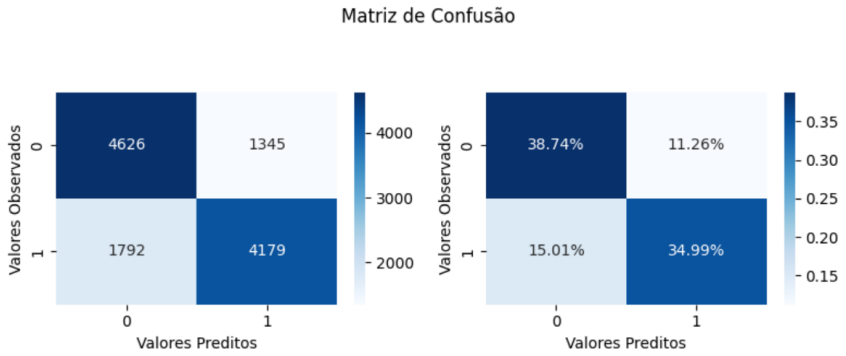

Figura 13 – Matriz de Confusão para a Floresta Aleatória

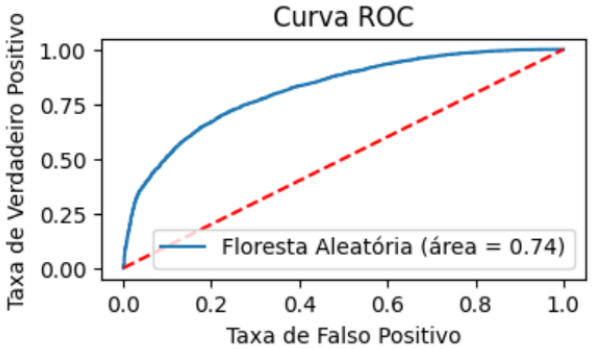

Figura 14 – Curva ROC para a Floresta Aleatória

Os principais valores de coeficientes da floresta aleatória podem ser encontrados na Tabela 5, na qual notamos que os principais fatores são o uso de suporte ventilatório, a internação em UTI, a idade e a saturação.

Tabela 5 – Importância dos Principais Coeficientes da Floresta Aleatória

Preditor	Categoria	Floresta Aleatória
Suporte Ventilatório	Sim	0,288623
UTI	Sim	0,146592
Idade		0,145515
Saturação $O_2 < 95\%$	Sim	0,060216
Tosse	Sim	0,024058
Outros Fatores de Risco	Sim	0,015646
Imunodepressão	Sim	0,013266
Doença Renal Crônica	Sim	0,011126
Sexo	Fem.	0,010633
Doença Neurológica Crônica	Sim	0,010049

3.4 Redes Neurais

Para criação da rede neural usou-se a biblioteca keras, fizemos uma otimização de hiperparâmetros no espaço paramétrico em que variou-se o número de neurônios em cada camada oculta de 2 a 50, e uma taxa de *dropout* de 0,1 a 0,9. Após 15 iterações com uso de 30 epocas em cada iteração, encontrou-se como melhor resultado usar 25 neurônios em cada camada oculta, e uma taxa de *dropout* de aproximadamente 0,26.

Na Figura 15 encontramos a matriz de confusão para a regressão logística, e na Figura 16 encontramos sua curva ROC.

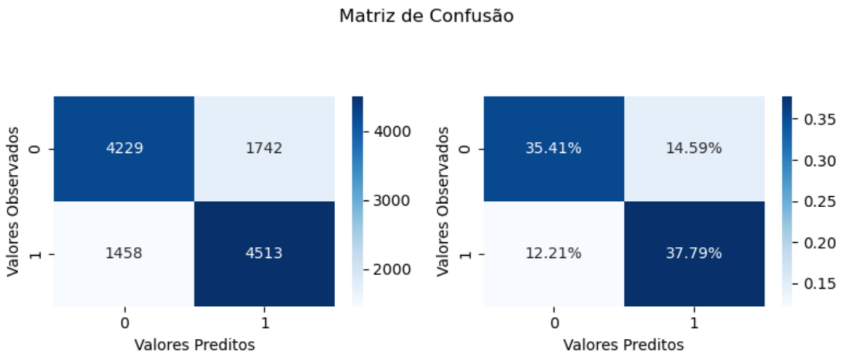

Figura 15 – Matriz de Confusão para a Rede Neural

A rede neural não possui coeficientes que possam ser interpretados.

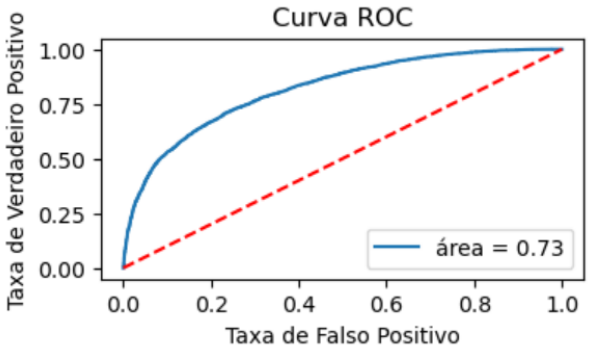

Figura 16 – Curva ROC para a Rede Neural

3.5 Discussão

Outros modelos de classificação foram testados, mas não houve nenhuma melhora significativa em relação aos modelos citados. Por isso, optou-se por focar apenas na regressão logística, na floresta aleatória e na rede neural, pois são três modelos diferentes, sendo a regressão logísitica um modelo bem conhecido e interpretável, a floresta aleatória que é um modelo conhecido por ter boa performance em dados tabulares e a rede neural pois é um modelo que tem ganhado muita notoriedade recentemente, embora não seja interpretável e seja mais computacionalmente intensivo.

As métricas de qualidade preditiva utilizadas e os valores obtidos para cada modelo podem ser encontrados na Tabela 6. Resultados similares foram encontrados por (SILVADARCY RISOMARIO; NETO, 2022), mas o presente trabalho se difere por explorar o uso de redes neurais, utilizar dados de 2022 e aplicar valores SHAP para ter outra forma de explicabilidade dos modelos.

Tabela 6 – Métricas de Avaliação dos Modelos

Modelo	Acurácia	Recall	Precision	f1 score	ROC-AUC
Regressão Logística	0,73	0,70	0,75	0,72	0,73
Floresta Aleatória	**0,74**	0,69	**0,76**	0,72	**0,74**
Rede Neural	**0,74**	**0,73**	0,74	**0,73**	**0,74**

Notamos, na Tabela 6, que as três técnicas utilizadas apresentam um poder preditivo muito similar, com a acurácia acima de 70%.

Para a regressão logística, podemos ver que as três variáveis com maiores valores SHAP são suporte ventilatório, idade e UTI. Estas três variáveis também figuram entre as mais importantes olhando para a razão de chances.

Para a floresta aleatória, podemos ver que as três variáveis com maiores valores SHAP são suporte ventilatório, idade e UTI. Essas três também são apontadas como as mais importantes de acordo com o método intrínseco da floresta aleatória.

Tabela 7 – Principais Valores SHAP para os Modelos

Variável	Categoria	Reg. Log.	Flor. Ale.	Rede Neural
Desconforto Respiratório	Sim	**0,1412**	0,0201	0,0217
Dispneia	Sim	0,1328	**0,0206**	0,0242
Idade		**0,5262**	**0,0477**	**0,0892**
Outros Fatores de Risco	Sim	0,1238	0,0127	0,0196
Recebeu Vacina COVID-19	Sim	0,1212	0,0039	0,0147
Região do Brasil	Sul	0,1360	0,0095	**0,0300**
	Sudeste	0,0679	0,0102	0,0118
	Centro Oeste	0,0657	0,0013	0,0141
	Nordeste	0,0339	0,0010	0,0095
Saturação $O_2 < 95\%$	Sim	0,1273	**0,0313**	0,0234
Suporte Ventilatório	Sim	**0,7294**	**0,0969**	**0,1188**
Tosse	Sim	**0,1989**	0,0196	**0,0291**
UTI	Sim	**0,3592**	**0,0629**	**0,0631**

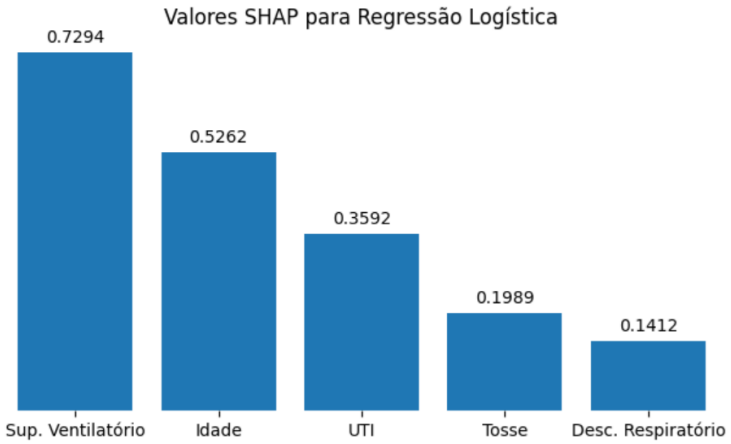

Figura 17 – Principais valores SHAP para regressão logística.

Por fim, para a rede neural, podemos ver que as três variáveis com maiores valores SHAP são suporte ventilatório, idade e UTI.

Do exposto, observa-se que as variáveis suporte ventilatório, idade e UTI são as mais importantes para todos os modelos. Assim, podemos comparar essa importância usando a mesma metodologia e termos mais certeza do resultado. Com isso, os profissionais da área podem saber em quais fatores deveriam prestar mais atenção a fim de entender como a evolução da doença pode se dar.

Ainda comparando os resultados das três técnicas, notamos que a Rede Neural identificou a Região Sul e a variável outra comorbidade como relevantes, enquanto que a Regressão Logística identificou a tosse e o desconforto respiratório e a Floresta encontrou a saturação e e dispneia. Desta forma, podemos utilizar essas técnicas de forma complementar, ajudando na identificação de fatores associados que passariam despercebidos caso utilizássemos uma única técnica.

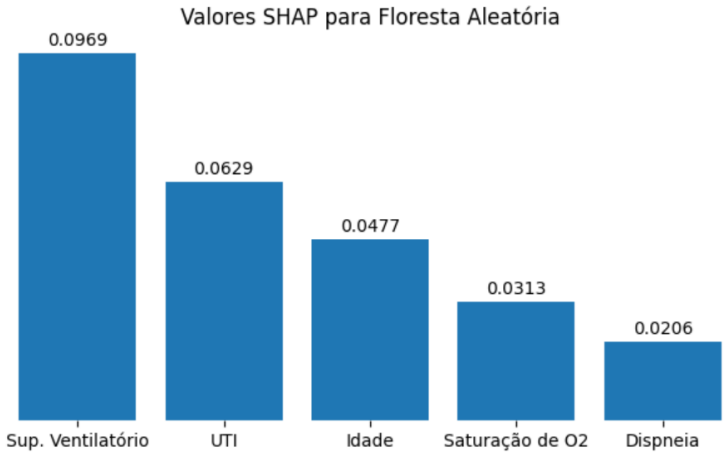

Figura 18 – Principais valores SHAP para floresta aleatória.

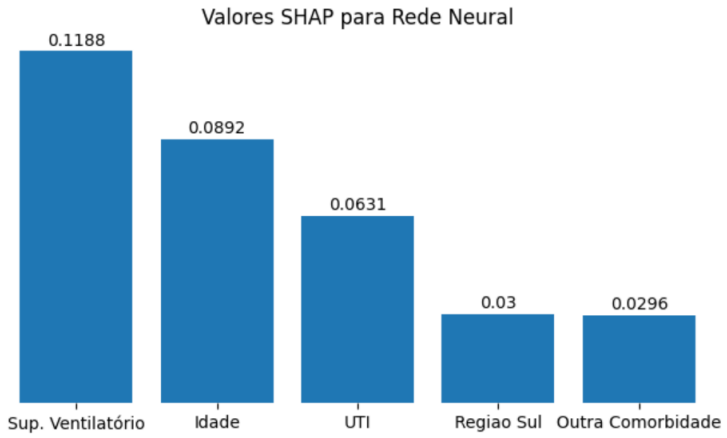

Figura 19 – Principais valores SHAP para rede neural.

Além disso, SHAP permite tanto a interpretabilidade global, ou seja, do modelo como um todo conforme Figura 20, quanto a interpretabilidade local, isto é, entender a importância e a influência de cada variável na previsão de cada paciente, como podemos observar na Figura 21.

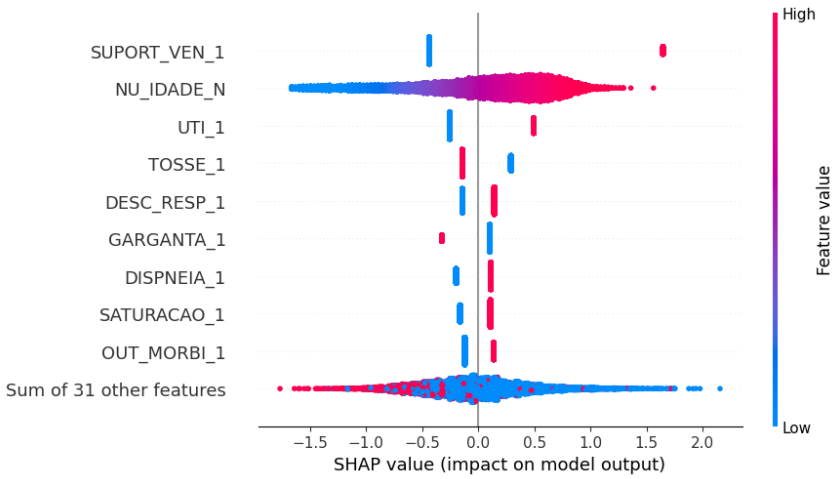

Figura 20 – Usando Valores SHAP para interpretabilidade global.

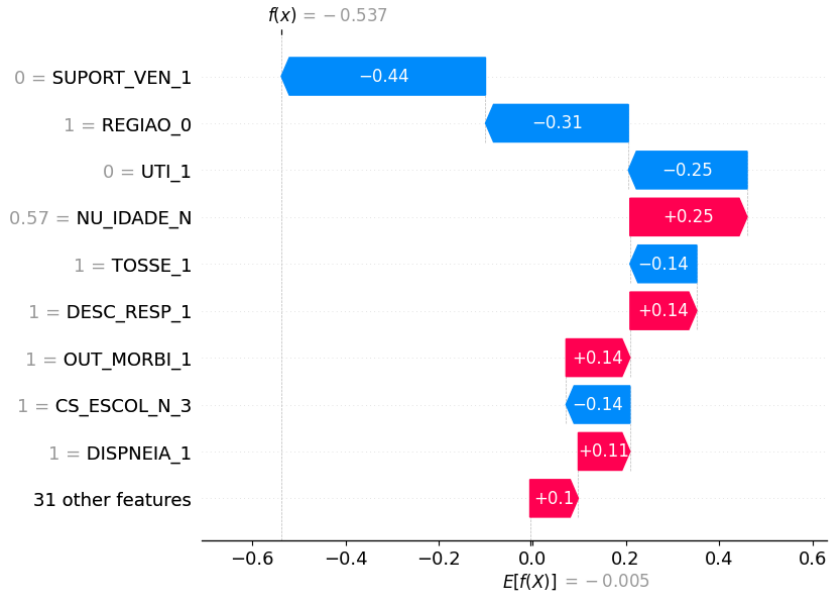

Figura 21 – Valores SHAP para entender a previsão de um paciente pela Regressão Logística.

Na Figura 21 temos a contribuição de cada variável para a previsão do desfecho de um paciente feita com a regressão logística. Trata-se de um paciente do sexo masculino, de 75 anos, da região norte, raça branca, escolaridade superior, que apresentou febre, tosse, dispnéia, desconforto respiratório, saturação de oxigênio, outra comorbidades e tomou vacina pra COVID e que não veio a óbito.

Na Figura 21, abaixo do eixo x, o valor de referência ($E[f(X)]$) é exibido, indicando o valor esperado do modelo avaliado no conjunto de dados. Os valores SHAP de cada variável são somados para alinhar a saída do modelo considerando todas as variáveis.

Cada barra no gráfico representa a contribuição de uma variável para a previsão final. As barras vermelhas indicam que o valor da variável "puxou" a previsão para cima, isto é, contribuiu positivamente para a classe prevista que no presente trabalho seria o óbito. As barras azuis indicam que a variável "puxou" a previsão para baixo, isto é, contribuiu negativamente para a classe prevista. O comprimento da barra representa a magnitude da contribuição. :

Ao lado de cada barra, o gráfico exibe o valor específico da variável que foi usado para aquela instância. No topo da cascata, o gráfico chega ao valor final da previsão. Este é o resultado do valor base ajustado pelas contribuições, positivas ou negativas, de cada variável. Esse valor final pode ser uma probabilidade em problemas de classificação ou um valor numérico em problemas de regressão.

Do exposto, podemos notar que, neste caso, se um dos três modelos testados tivesse que ser escolhido para ser utilizado, poderíamos seguir com a regressão logística, pois é um modelo mais bem conhecido, interpretável e não é computacionalmente intensivo.

Capítulo 4

Considerações Finais

Especialmente no caso da COVID-19, por ser uma doença com características não totalmente conhecidas e que tinha rápida evolução, sistemas ou métodos que consigam determinar quais são os pacientes mais suscetíveis a desenvolverem um quadro grave da doença são mais ferramentas para ajudar os profissionais de saúde a tomarem decisão no momento da triagem.

Em especial, ao entendermos quais as principais variáveis que modelos com boa capacidade preditiva usam, os profissionais de saúde podem prestar mais atenção a estas características. Nesse sentido, além de alguns métodos intrínsecos de alguns modelos clássicos, podemos usar valores SHAP para termos maior interpretabilidade ou para entender melhor modelos do tipo "caixa preta" como os que estão ficando cada vez mais comuns.

No momento da triagem uma pessoa pode apresentar um bom quadro clínico, mas se modelos apontarem que estas pessoas tem maiores chances de desenvolverem quadros graves da doença com o tempo, isso é mais um insumo para o profissional de saúde decidir como tratar a pessoa em questão.

Especialmente nos casos em que a evolução da doença é rápida, conhecer pessoas com fatores ligados a uma possível piora da doença pode ser decisivo.

Para uma sequência ou aprimoramente deste trabalho, pode-se considerar os seguintes pontos:

- mudar a categorização de algumas covariáveis (idade, uso de suporte ventilatório, entre outros) a fim de obter um maior poder preditivo;

- avaliar o efeito dos fatores associados na probabilidade de óbito dos pacientes;

- utilizar outras técnicas de aprendizado supervisionado.

Referências

BAGABIR, e. a. Covid-19 and artificial intelligence: Genome sequencing, drug development and vaccine discovery. *Journal of Infection and Public Health*, v. 15, p. 289–296, 2022.

BASTIAN, M. *GPT-4 has more than a trillion parameters - Report*. [S.l.], 2023. Disponível em: <https://the-decoder.com/gpt-4-has-a-trillion-parameters/>. Acesso em: 06 jun. 2023.

BERKSON, J. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, n. 39, p. 357–365, 1944.

COMITO CARMELA; PIZZUTI, C. Artificial intelligence for forecasting and diagnosing covid-19 pandemic. *Artificial Intelligence in Medicine*, v. 128, n. 102286, 2022.

DABBAGH, R. e. a. Harnessing machine learning in early covid-19 detection and prognosis: A comprehensive systematic review. *Cureus*, v. 15, 2023.

GUDIGAR, e. a. Role of artificial intelligence in covid-19 detection. *Sensors*, v. 21, n. 8045, 2021.

HO, T. K. Random decision forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, p. 278–282, 1995.

HUANG C.; WANG, Y. e. a. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, v. 395, n. 10223, p. 497–506, 2020.

HUANG, e. a. Artificial intelligence in the diagnosis of covid-19. *International Journal of Biological Sciences*, v. 17, n. 6, p. 1581–1587, 2021.

LUNDBERG S. M.; LEE, S.-I. *A unified approach to interpreting model predictions*. [S.l.], 2017. Disponível em: <https://dl.acm.org/doi/10.5555/3295222.3295230>. Acesso em: 23 jun. 2024.

MCCULLOCH, W. W. P. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943.

SILVADARCY RISOMARIO; NETO, R. d. S. Inteligência artificial e previsão de óbito por covid-19 no brasil: uma análise comparativa entre os algoritmos logistic regression, decision tree e random forest. *Saúde debate*, v. 46, 2022.

VERHULST, P.-F. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique*, n. 10, p. 113–121, 1838.

VERHULST, P.-F. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, n. 18, 1845.

ZAERI, N. Artificial intelligence and machine learning responses to covid-19 related inquiries. *Journal of Medical Engineering and Technology*, v. 6, p. 301–320, 2024.