Análise Estatística de Casos de Síndrome Respiratória Aguda Grave no SUS

Rodrigo Abreu
Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
raafm@cin.ufpe.br

João Pedro Ribeiro
Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
jprsd@cin.ufpe.br

Rayhene Ranuzia
Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
rrda@cin.ufpe.br

Fernando Macedo
Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
frpmn@cin.ufpe.br

João Victor
Centro de Informática
Universidade Federal de Pernambuco
Recife, Pernambuco
jvol@cin.ufpe.br

***Abstract*—A pandemia de COVID-19 dos últimos anos causou vários casos de internação por complicações no sistema respiratório. Esse trabalho visa analisar estatisticamente os dados de internações por Síndrome Respiratória Aguda Grave (SRAG) disponibilizados pelo governo brasileiro. Além disso, será proposto um esquema de limpeza e recuperação de dados faltantes por meio de modelos de Machine Learning.**

***Index Terms*—SUS, SRAG, COVID-19, Machine Learning, Naive Bayes, classificador, paciente.**

I. Introdução

COVID-19 (do inglês: Coronavirus Disease 2019, em português: Doença por Coronavírus – 2019) é uma doença infeciosa causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). Cerca de 80% das infeções pelo SARS-CoV-2 confirmadas têm sintomas ligeiros de COVID-19 ou são assintomáticos, e a maioria recupera-se sem sequelas. No entanto, 15% das infeções resultam em COVID-19 severa com necessidade de oxigênio e 5% são infeções muito graves que necessitam de ventilação assistida em ambiente hospitalar. Os casos mais graves podem evoluir para pneumonia grave com insuficiência respiratória grave, septicemia, falência de vários órgãos e morte [2].

O Ministério da Saúde (MS), por meio da Secretaria de Vigilância em Saúde (SVS), desenvolve a vigilância da Síndrome Respiratória Aguda Grave (SRAG) no Brasil, fortalecida desde a pandemia de Influenza A(H1N1).

Recentemente (2020), a vigilância da COVID-19, a infecção humana causada pelo SARS-CoV-2, foi incorporada na rede de vigilância da Influenza e outros vírus respiratórios.

Um conjunto de tabelas derivadas da vigilância da SRAG foi disponibilizada no site do opendatasus, podendo ser baixada em formato CSV. Os dados usados nesse trabalho são provenientes das tabelas de 2020 [4] , 2021 e 2022 [3]. Como as tabelas são atualizadas semanalmente, a versão usada nesse trabalho pode divergir dos dados mais recentes.

II. Objetivos

O presente trabalho tem como objetivo fazer uma análise estatística dos dados de SRAG apresentados pelo SUS entre 2020 e o início de 2020, propor uma heurística para limpeza e recuperação dos dados. Por fim, proporemos um classificador, capaz de determinar se uma pessoa infectada com covid terá uma evolução para óbito ou não.

III. Base de dados

3 bases de dados foram utilizadas: os casos de SRAG de 2020, de 2021 e de 2022, (todas disponibilizadas pelo opendatasus).

IV. Análise exploratória dos dados

A. Idade

Para evitar casos com idades improváveis, apenas consideramos os que possuíam idade entre 0 e 100 anos. Porém a base possuía idades até 150 anos e também negativas.

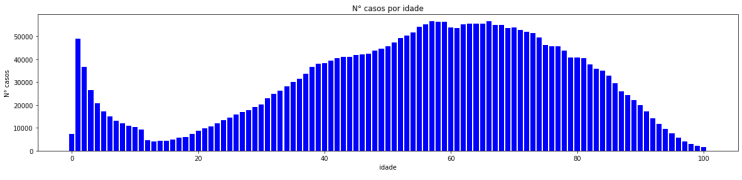

Fig. 1. Número de casos de SRAG por idade

B. Comorbidades

Distribuição de pessoas marcadas com fator de risco pelo SUS (coluna FATOR_RISC = 1).

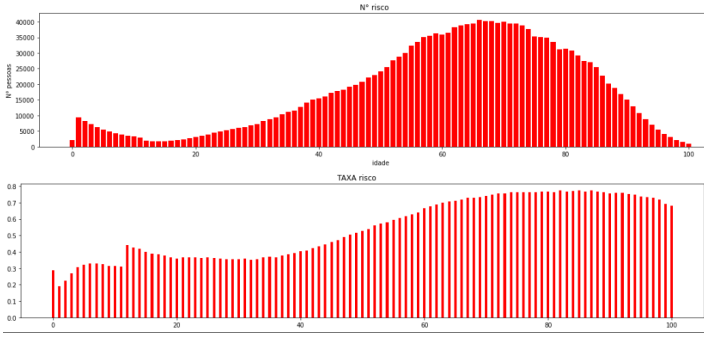

Fig. 2. Número de casos de SRAG em pessoas com comorbidades por idade

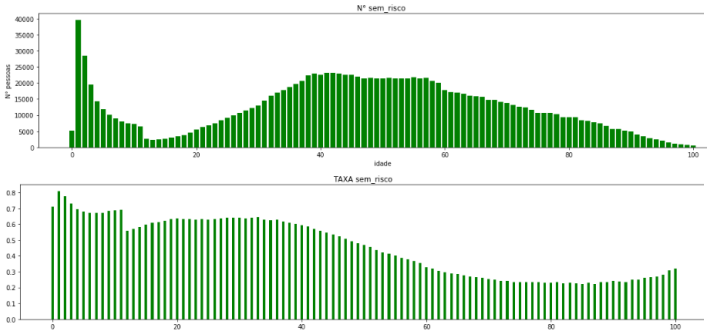

Fig. 3. Número de casos de SRAG em pessoas sem comorbidades por idade

1) Distribuições por idade: Vermelho, com comorbidades, [figura 2].

Média de idade: 61.3

desvio padrão da idade: 19.86

faixa de idades dentro de 1 desvio padrão: [41.44, 81.16]

faixa de idades dentro de 2 desvios padrões: [21.58, 101.02]

Verde, sem comorbidades, [figura 3].

Média de idade: 45.96

Desvio padrão da idade: 23.6

idades dentro de 1 desvio padrão: [22.36, 69.56]

idades dentro de 2 desvios padrões: [-1.25, 93.16]

2) risco X risco grave: Nesta parte do nosso projeto analisamos como os fatores de risco classificados pelo o SUS e os fatores de risco graves, aqueles que foram selecionados por nós como, cardiopatia, imunossupressão, obesidade mórbida e entre outros, além da idade ser levada em consideração também como uma agravante, estavam ligados diretamente a quantidade de casos em nossa base, além do fato de que a porcentagem de letalidade também é maior nessa nova classe que criamos[figura 4 e 5].

3) Distribuição de IMC: Nesta sessão do analisamos o IMC das pessoas cadastradas no nosso banco de dados e nos deparamos com situações atípicas, pois haviam pessoas com IMC 0, que se trata de um caso impossível e pessoas com o IMC de 3000, outro caso impossível, portanto para levarmos em consideração apenas casos reais, estabelecemos o nosso IMC entre 10 e 80, pois taxa de IMC acima de 40, representam obesidade mórbida a qual era aquela que estávamos interessados[figura 6].

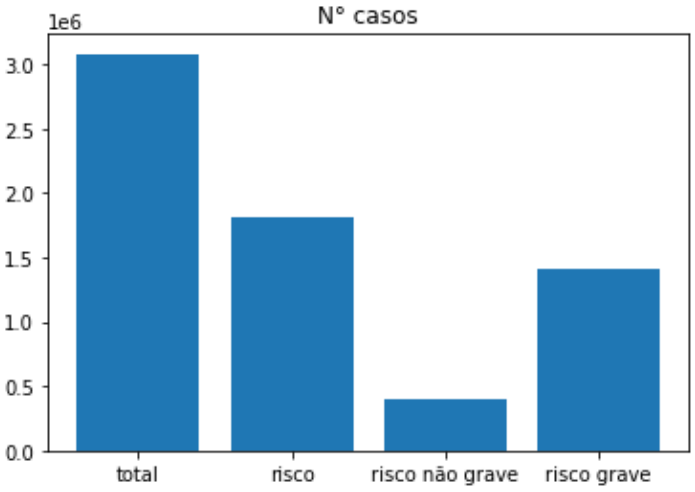

Fig. 4. Número de casos visto de uma perspectiva de risco X risco alto

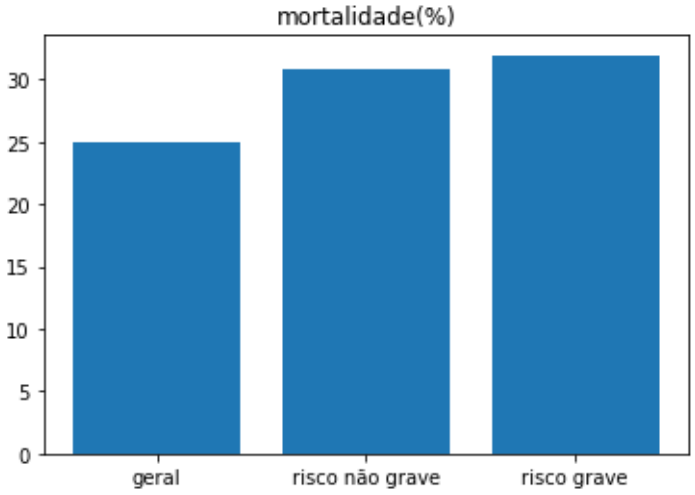

Fig. 5. Número de casos visto de uma perspectiva de risco X risco alto

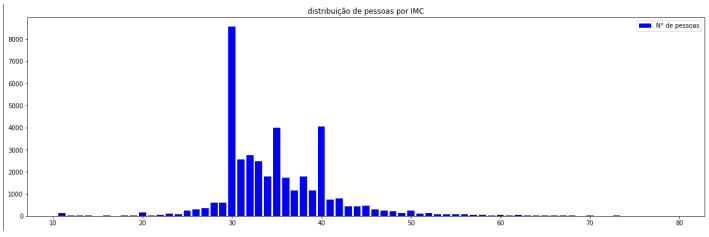

Fig. 6. Distribuição de pessoas por IMC

C. Vacinação

1) taxa de vacinação: Para a vacinação utilizamos duas métricas diferentes: a taxa e a distribuição. A distribuição se refere ao total de doses que foram aplicadas, sendo distribuídas por idade. A taxa refere-se a porcentagem do total de pessoas por idade que receberam as doses da vacina [figure 7].

Média de idade: 65.02

Desvio padrão da idade: 18.27

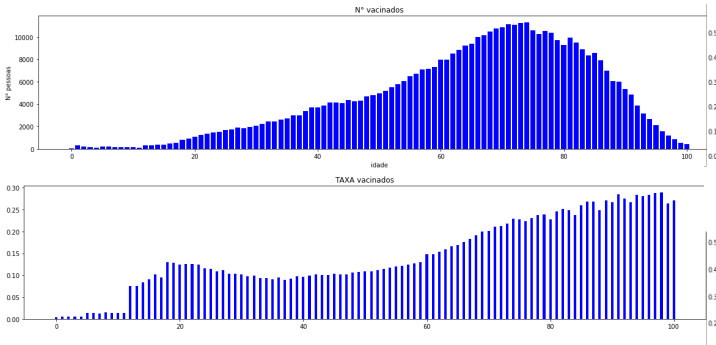

Fig. 7. Distribuição e taxa de vacinação por idade

idades dentro de 1 desvio padrão: [46.75, 83.29]

idades dentro de 2 desvios padrões: [28.47, 101.56]

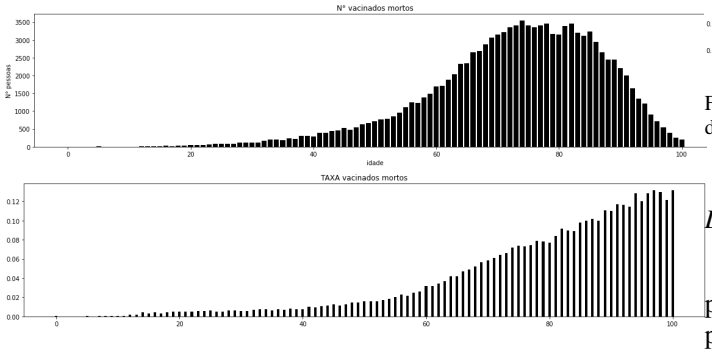

Fig. 8. Distribuição e taxa de não vacinados mortos por idade

2) taxa de vacinados mortos: Como a vacina foi liberada em ordem de idade , sendo os mais velhos ordinariamente os primeiros a receber, temos uma média de idade dos vacinados em uma faixa etária de idosos.

A taxa aumenta até um valor aproximado de 30% de vacinados entre os casos com pacientes mais velhos. Isto era esperado, uma vez que a base leva em consideração casos de 2020 até 2022 e em grande parte desse período a vacina não estava liberada para população. A análise do gráfico, portanto, não deve induzir que a taxa de vacinação está baixa no presente momento.

Podemos notar que a mortalidade mostrada na figura 8 (máximo entre os mais velhos em cerca de 12%) é muito menor que a média da população (cerca de 25%). Isso ocorre porque a figura 8 foi obtida dividindo o número de pessoas que tomaram a vacina e morreram de uma certa idade pelo número total de pessoas desta idade. Como a base possui muitos casos não vacinados, a tendência de diminuir a mortalidade por falta de pessoas (vacinados e mortas).

Uma análise mais coerente foi feita dividindo o número de pessoas que tomaram a vacina e morreram de uma certa idade pelo número total de pessoas vacinadas desta idade. A figura 9 mostra está relação. Vemos que a melhor mudança ocorre para indivíduos com idade superior a 60 anos.

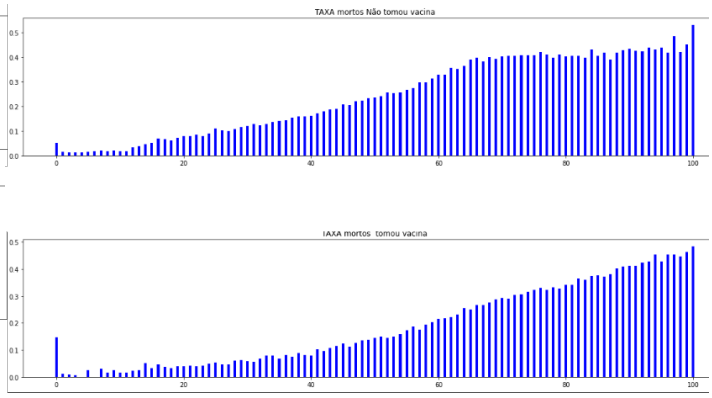

Fig. 9. Taxa de mortalidade de vacinados por idade ente vacinados.

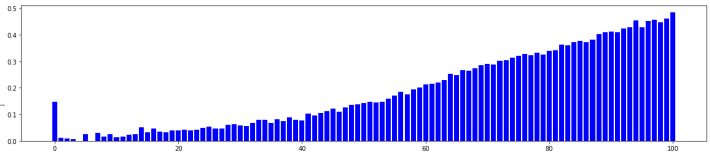

Fig. 10. Taxa de mortalidade apenas para vacinados: vacinados que morreram dividido total de vacinados por idade.

D. Inferência das variantes

As variantes de SARS-COV-2 que surgiram ao longo da pandemia, ocasionaram aumentos de casos em momentos específicos, como podemos ver na figura abaixo que a distribuição diária de casos possui ondas de casos.

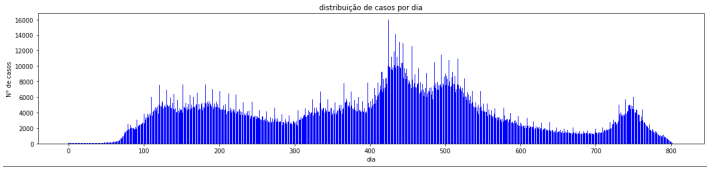

Fig. 11. N° de casos diários

A variante que predominava durante a data dos primeiros sintomas do paciente foi estimada visualmente a partir dos gráficos dos números de casos por dia, que tiveram a imagem modificada para melhor visualização.

A distribuição do número de casos (considerando o dia dos primeiros sintomas como o dia do caso) por dia pode ser representada por uma função $f : [0, 800] -> [0, 13.000]$, em que $[0, 800]$ é a contém os dias possíveis na base e o máximo de casos no dia foi 13.000.

A seguinte sequência de operações foi aplicada na função:

1) operação linear de normalização (subtrair da média e dividir por uma constante) na função. Sua imagem ir de $[0, 13.000]$ a $[-1, 1]$. Chamemos esta função normalizada de f_{norm}.
2) Transformada de Fourier por meio do algoritmo Fast Fourier Transform (FFT) [1].

3) Filtragem dos maiores valores no *power spectrum* gerado como output da FFT. Ondas com valor inferior a determinado valor foram descartadas.
4) Transformada inversa de fourier usando o inverse fast fourier transform (IFFT). Consequentemente, recuperando uma outra função com formato semelhante ao de f_{norm}, mas visualmente mais simples para identificação das ondas de cada variante. Chamemos a função obtida ao final deste passo de f_filt.

Estas operações foram aplicadas em todos os casos, pois consideramos que a principal causa dos aumentos repentinos do número de casos de SRAG foram consequência do surgimento das vairantes.

Note que a função f_{norm} é análoga a um sinal que tem o ruído filtrado pela sequência de operações descrita anteriormente, e f_filt o sinal limpo, veja a figura a seguir:

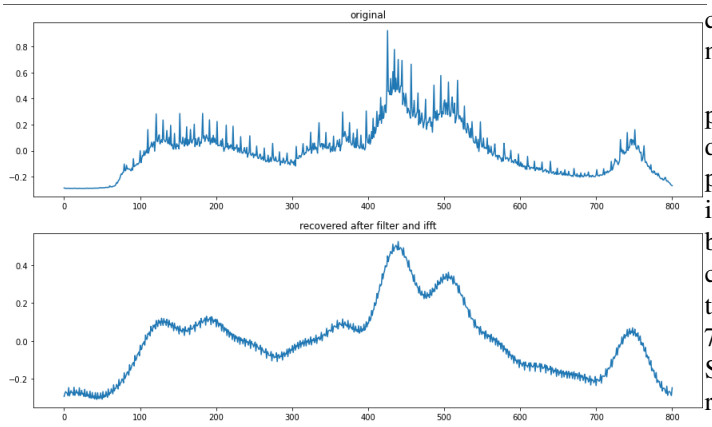

Fig. 12. f_{norm}(cima) e f_filt(baixo)

Na imagem abaixo, podemos ver a onda da variante predominante no fim de 2021 e começo de 2022 (ômicron):

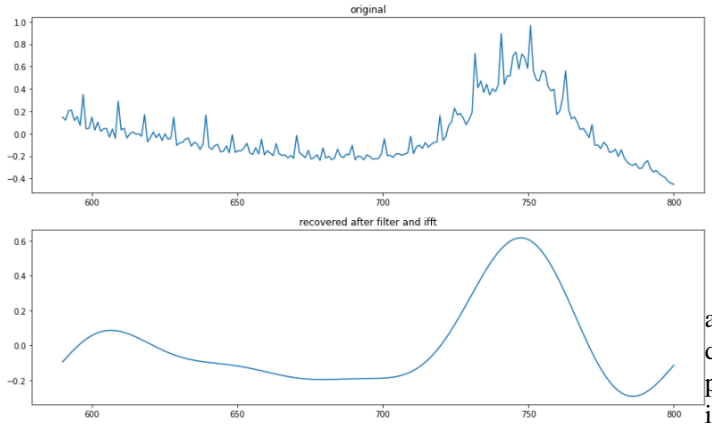

Fig. 13. f_{norm}(cima) e f_filt(baixo)

V. Modelos de machine learning

A. Classificador Ingênuo de Bayes

1) Implementação: Após tratar os dados e categorizar todos os dados contínuos que tínhamos, fizemos uma implementação do classificador ingênuo de Bayes categórico para testa-lo usando a nossa base. A implementação foi feita usando os conceitos básicos de calcular a probabilidade de um evento acontecer dado que outro aconteceu. No nosso caso, calculamos a probabilidade de um paciente morrer ou sobreviver, e multiplicamos pela probabilidade de ele morrer ou sobreviver, dados os parâmetros selecionados como mais importantes na análise dos dados, que foram a faixa de idade em que se encontravam, a onda de COVID na época em que o paciente foi internado, uma classificação de quanto tempo fazia desde a ultima dose tomada da vacina, se o paciente teria ido para a UTI e se possuía alguma comorbidade.

2) Testes: Foi então usada uma fração da base de dados para treinar o classificador e esses mesmos dados foram usados para calcular a acurácia do mesmo. Após essa classificação, foi então usada uma implementação já pronta desse mesmo classificador, da biblioteca SKLearn, para então compararmos com o modelo que fizemos do zero. Feita essa comparação nós chegamos aos seguintes resultados:

3) Comparando os resultados: Obtivemos Resultados muito parecidos entre os do classificador que implementamos e os do classificador da biblioteca SKLearn, usamos uma parte pequena da base de dados para os testes pois não foi uma implementação tão eficiente temporalmente, com 0.1% da base de dados obtivemos uma acurácia de 73.76% no nosso classificador e 73.28% no classificador do SKLearn, tentamos também com 1% da base de dados e obtivemos resultados de 73.35% de acurácia no nosso classificador e de 72.04% no do SKLearn. Podemos observar então que os modelos chegam a resultados próximos, e em alguns casos, superiores no modelo que implementamos,

B. Redes neurais

Foram feitas 4 redes neurais:

1) sem informações sobre o final do caso, classifica a evolução (se sobreviverá ou não).
2) com informações sobre o final do caso (tempo passado na UTI e a Classificação final registrada), classifica a evolução (se sobreviverá ou não).
3) com informações sobre o final do caso(tempo passado na UTI e a evolução, ou seja, se sobreviveu ou não), classifica o caso como COVID-19 ou não .
4) sem informações sobre o final do caso, classifica o caso como COVID-19 ou não .

1) acurácia: As redes para prever a sobrevivência obtiveram acurácia dentro da faixa: 75-82 %, enquanto para classificação de Covid-19 a acurácia de ambos foi aproximadamente 90% para ambos os classificadores. Isso pode ser explicado pela importância das features, que será discutida posteriormente.

2) modelo: As redes são do mesmo modelo: 4 layers com activation function do tipo Relu entre cada layer e aplicação de softmax no último layer com 2 outputs (se sobreviveu ou não nas duas primeiras redes, e se foi caso de covid ou não para segunda rede), a resposta de classificação da rede seria aquela output com maior valor. Foram testados modelos de tamanhos diferentes (5 layers, 6 layers, mais ou

menos neurônios por layer), mas nenhum obteve resultado significativamente melhor para citarmos aqui. O otimizador usado foi o Adam e a biblioteca usada para implementação foi a pytorch.

3) Experimentos: Além de variar o formato da rede pela profundiadde e número de neurônios, tentou-se melhorar a qualidade da rede com o modo de equilibrar os casos. A base, após ser limpa, possuía a grande maioria dos casos como classificados com Covid-19, em proporção suficiente para causar esquecimento catastrófico dos outros casos durante o treinamento, i.e., a rede só aprenderia apropriadamente pessoas com Covid-19.

O primeiro modo de equilibrar foi realizando o treino com todos os casos sem covid-19 e uma amostra dos casos de Covid-19 de tamanho mais próximo. O segundo modo foi por repetição dos casos de sem covid. Metade dos casos sem Covid-19 foram escolhidos randomicamente para treino , em seguida foi repetida várias vezes no mesmo dataset e misturada com igual quantidade de casos com covid-19.

O segundo modo gerou resultados ligeiramente melhores (aumento de menos de 5% na acurácia), provavelmente pela maior disponibilidade de dados para treinamento.

Tentou-se também usar apenas a colunas mais genéricas, como: fator de risco , comorbidades, etc. E em outra situação usar as colunas especificas de cada comorbidade como: asma, diabetes, cardiopatia, etc. A diferença de acurácia após os dois treinos foi desprezível.

Por fim, aumentando o número de iterações e de samples no treino gerou o mesmo efeito em relação a treinos menores (em tempo e quantidade de dados): a acurácia aumentava e a loss diminuía, porém a melhoria do modelo aumentava de modo imensamente desproporcional a quantidade de tempo e samples empregados. Portanto, pode-se dizer que o atual treinamento (200 iterações e cerca de 1M de casos) prescinde de mais casos ou de mais tempo, tendo alcançado um limite próprio.

4) Uso: As redes de previsão de evolução final não obtiveram resultado satisfatório, porém as de classificação de covid sim. Portanto podemos usar as redes 3 e 4 para preencher os campos faltantes (de CLASS_FIN) na base .

C. Outros modelos

Com o objetivo de melhorar a nossa acurácia,recorremos as técnicas de machine learning (ML),para isso utilizamos a biblioteca scikit-learning, que já possui internamente os classificadores que iremos utilizar implementados e também a biblioteca XGBoost, pois ela fornece uma estrutura interna capaz de aumentar os gradientes, além disso também foram usadas duas métricas diferentes para a acurácia, accuracy score e cross value score.

1) Modelos de classificador: Foram utilizados 5 classificadores diferentes:

1) Regressão Logística.
2) Árvore de decisão.
3) Floresta Randômica.
4) Kneighbor.
5) XGBClassifier.

VI. Experimentos

1) Desempenho de cada classificador:

model	accuracy score	cross value score
Regressão Logística	86.85%	86.85%
Árvore de decisão	97.54%	88.43%
Floresta Randômica	97.54%.	90.16%
Kneighbor	93.05%	90.52%
XGBClassifier	91.30%	91.19%

2) Importância das features: Medimos a importância das features no regressor logístico e na árvore de decisão.

No regressor logístico observamos o valor do módulo dos coeficientes, pois grandes valores podem indicar maior importância da feature correspondente, uma vez que seu valor terá maior influência no resultado da rede. Como as variáveis foram normalizadas para uma mesma faixa de valores, esta premissa foi considerada verdadeira .

Podemos ver que as principais features para inferência da sobrevivência são em ordem: a classificação de Covid-19 confirmada por exame, seguida da idade (ambas muito próximas), seguidas do tipo de variante no período que ocorreram os primeiros sintomas, informações da vacina da covid, fator de risco grave, etc

feature	coefficient
CLASSI_COV_EXAME	4.290000
IDADE	4.210000
ONDA	-1.800000
TEMP_ULT_DOSE	1.510000
VACINA_COV	1.100000
FATOR_RISC_GRAVE	-0.780000
VACINA	-0.640000
FATOR_RISC	-0.550000
COMORB_GRAVE	0.200000
UTI	0.140000

Usando o modelo de árvore de decisão disponibilizado pelo sklearn, podemos usar o atributo *feature_importances_* para determinar a importância.

Note que a ordem das 3 primeiras features é a mesma que a anterior.

feature	importance
CLASSI_COV_EXAME	0.578
IDADE	0.2
ONDA	0.104
TEMP_ULT_DOSE	0.035
VACINA	0.025
UTI	0.024
COMORB_GRAVE	0.014
VACINA_COV	0.012
FATOR_RISC	0.005
FATOR_RISC_GRAVE	0.004

VII. Conclusões

Após testar os diferentes modelos, podemos notar que o Kneighbor e a floresta randômica são os melhores algoritmos para o nosso objetivo, de acordo com a accuracy score e o cross value score, também podemos vê que uma boa limpeza em base de dados sempre se faz necessário, pelo fato de que podem haver dados corrompidos.

References

[1] James W Cooley and John W Tukey. "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of computation* 19.90 (1965), pp. 297–301.

[2] Wikipedia. "COVID-19". In: *https://pt.wikipedia.org/wiki/COVID-19*. 2020.

[3] Wikipedia. "OPENDATA SUS 20". In: *https://opendatasus.saude.gov.br/dataset/srag-2020*.

[4] Wikipedia. "OPENDATA SUS 21 e 22". In: *https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022*.