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 infecciosa causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). Cerca de 80% das infecções 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 infecções resultam em COVID-19 severa com necessidade de oxigênio e 5% são infecções 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 (trecho de [3]).

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 disponibilizado (em formato CSV) no site https://opendatasus.saude.gov.br/. Os dados usados nesse trabalho são provenientes das tabelas de 2020 [5] , 2021 e 2022 [4]. 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, propomos um classificador, capaz de determinar se uma pessoa infectada com COVID-19 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). As colunas escolhidas para análise foram:

- NU_IDADE_N (idade),
- fatores de risco: FATOR_RISC, CARDIOPATI, ASMA, PNEUMOPATI, DIABETES, IMUNODEPRE, OBESIDADE, OBES_IMC,
- informações da vacinação e datas/semanas epidemiológicas: VACINA_COV, VACINA, DOSE_1_COV, DOSE_2_COV, DOSE_REF, DT_SIN_PRI, DT_NOTIFIC, DT_NASC, SEM_NOT, SEM_PRI, DT_UT_DOSE, DT_INTERNA, DT_ENTUTI, DT_SAIDUTI, DT_EVOLUCA,
- informações do caso e exames: UTI (se foi para UTI), EVOLUCAO (cura ou óbito), CLASSI_FIN (Diagnóstico final do caso), TOMO_RES (Aspecto Tomografia), RES_AN (Resultado do Teste Antigênico), PCR_RESUL, PCR_SARS2, POS_AN_FLU, RES_IGG, RES_IGM, RES_IGA,

Para melhor descrição, consulte o dicionário de dados disponível na mesma página do opendatasus que disponibiliza as tabelas.

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 (observe tabela abaixo).

parâmetro	valor
mean	55.029
std	22.7997
min	-9.0
The 25% percentile	41.0
The 50% precentile	58.0
The 75% precentile	72.0
max	150.0

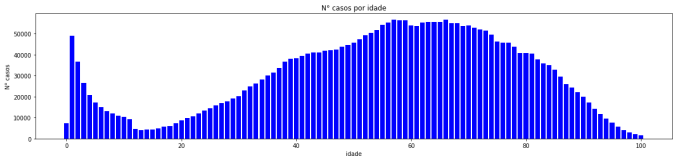

Fig. 1. Número de casos de SRAG por idade (apenas idades entre 0 e 100 anos). O eixo y possui marcado os números: 0, 10000, 20000, 30000, 40000, 50000; e o eixo x: de 0 até 100 (mostrando a cada 20).

B. Comorbidades

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

1) Distribuições por idade: Segue abaixo distribuições de número de casos com comorbidades por idade.

Em vermelho, com comorbidades Fig. 2:

parâmetro	valor
Média de idade	61.3
desvio padrão da idade	19.86
faixa dentro de 1 desvio padrão	[41.44, 81.16]
faixa dentro de 2 desvios padrões	[21.58, 101.02]

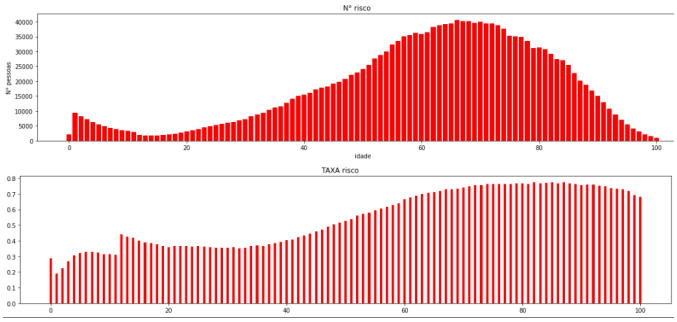

Fig. 2. Gráfico de cima: Número de casos com comorbidades por idade. Gráfico de baixo: taxa de casos com comorbidades por idade. No gráfico de cima, as marcações do eixo y começam em 0 , aumentam de 5000 em 5000. As taxas (gráfico de baixo) possuem marcações de 10% em 10% (começando em 0)

Em verde, sem comorbidades, Fig. 3

parâmetro	valor
Média de idade	45.96
desvio padrão da idade	23.6
faixa dentro de 1 desvio padrão	[22.36, 69.56]
faixa dentro de 2 desvios padrões	[-1.25, 93.16]

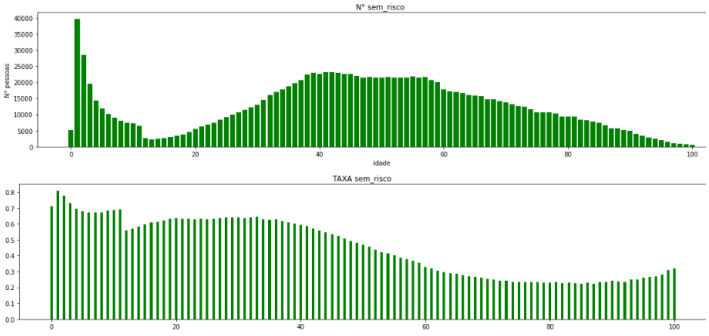

Fig. 3. Gráfico de cima: Número de casos sem comorbidades por idade. Gráfico de baixo: taxa de casos sem comorbidades por idade. No gráfico de cima, as marcaçoes do eixo y começam em 0 , e aumentam de 5000 em 5000. (mostrando a cada 20)

2) risco X risco grave: Analisamos os fatores de risco classificados pelo o SUS e alguns específicos como: cardiopatia, imunossupressão, obesidade mórbida, entre outros. Chamaremos estes de ” fatores graves”. A idade também foi considerada como fator de risco (e também fator grave). Constatou-se que, apesar da maior parte dos casos de risco serem graves (fig. 4) a mortalidade de pacientes com fator grave é semelhante aos de não graves (fig. 5).

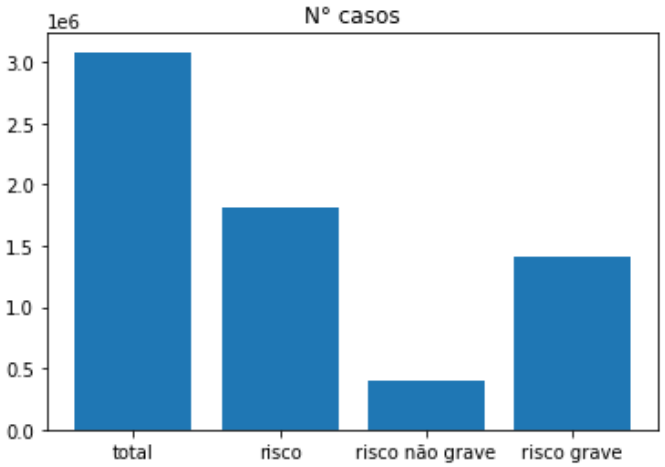

Fig. 4. Número de casos de acordo com o risco

3) Distribuição de IMC: Há situações atípicas, como: IMC 0 e IMC 3000, que se tratam de casos impossíveis, portanto estabelecemos o nosso IMC entre 10 e 80. A fig. 6 mostra uma distribuição de número de casos por IMC. Não sabemos com certeza a causa da distribuição de pessoas por IMC ser tendenciosa para mais de 30 (maior barra). Supomos seja pela não medição do IMC nos casos em que o paciente não apresentasse aparência física de obeso, evitando esforços desnecessários durante o atendimento.

4) Influência das comorbidades: Por meio de análise de feature importance (será explicado posteriormente), foram feitas 2 tabelas com a ordem de importância relativa de cada comorbidade escolhida da base (cardiopatia, diabetes,

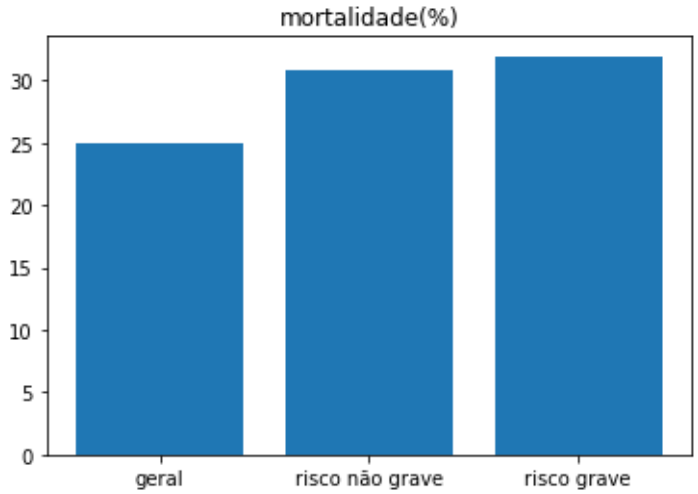

Fig. 5. Mortalidade de acordo com o risco

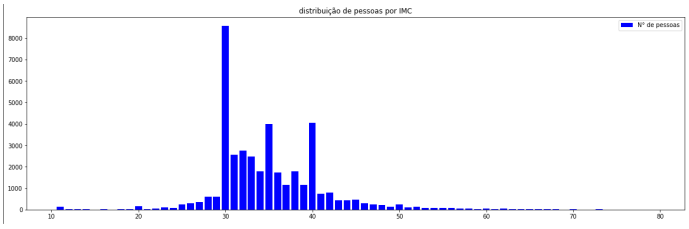

Fig. 6. Distribuição de pessoas por IMC. As marcações vão de 10 até 80, sendo a maior barra (maior número de casos) no IMC 30.

obesidade, pneumopatia, imunodeficiência , asma) e outras features para inferência da evolução do caso (sobrevivência ou não do paciente). A primeira tabela foi obtida por meio de um modelo de árvore de decisão com acurácia de 98,5%. A segunda tabela foi obtida por meio de um modelo de regressão logística com acurácia de 87,6%. No modelo de regressão logística a importância de cada feature foi estimada de acordo com o valor dos coeficientes do modelo após o treino (que podem ser negativos), isto será explicado posteriormente. Considere que quanto maior o módulo do coefficient maior a importância da feature.

A feature comorbidade corresponde a coluna FATOR_RISC fornecida pela base de dados.

A última tabela mostra a importância do maior para menor de acordo com cada modelo.

tabela de feature importance feita a partir de modelo de árvore de decisão

feature	importance
EXAME POSITIVO COVID	0.561
IDADE	0.192
ONDA	0.096
tempo desde ultima dose contra COVID	0.03
UTI	0.022
VACINA CONTRA GRIPE	0.021
CARDIOPATI	0.014
DIABETES	0.012
OBESIDADE	0.012
VACINA CONTRA COVID	0.012
COMORBIDADES	0.009
PNEUMOPATI	0.007
IMUNODEPRE	0.007
ASMA	0.005
OBESIDADE MORBIDA	0.001

tabela de feature importance feita a partir de modelo de regressão logística

feature	coefficient
EXAME POSITIVO COVID	4.3004
IDADE	3.5986
ONDA	-1.8482
OBESIDADE	1.6843
tempo desde última dose contra COVID	1.3821
OBESIDADE MORBIDA	1.0042
PNEUMOPATI	-0.9573
VACINA CONTRA COVID	0.9539
VACINA CONTRA GRIPE	-0.6684
COMORBIDADES	-0.538
IMUNODEPRE	-0.4457
DIABETES	0.3706
ASMA	-0.2527
CARDIOPATI	-0.1853
UTI	0.1269

tabela de feature importance, da maior para menor, de acordo com cada modelo

árvore de decisão	modelo de regressão logística
CARDIOPATIA	OBESIDADE
DIABETES	OBESIDADE MORBIDA
OBESIDADE	PNEUMOPATIA
COMORBIDADES	COMORBIDADES
PNEUMOPATIA	Imunodeficiência
Imunodeficiência	DIABETES
ASMA	ASMA
OBESIDADE MORBIDA	CARDIOPATIA

C. Classificação final

Para Fig. 7:

classificação final	quantidade de casos
influenza	18890
outro vírus respiratório	26120
outro agente etiológico	9063
não especificado	843998
COVID-19	1988453

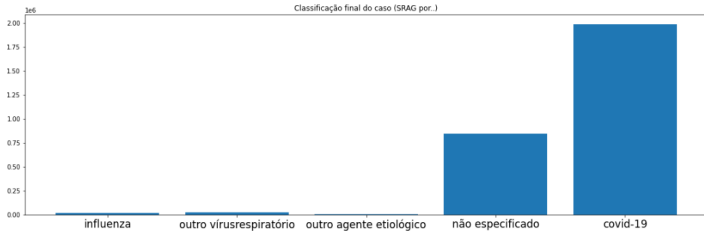

Fig. 7. quantidade de casos para cada classificação final (após limpeza da base)

Criamos uma coluna de classificação de COVID confirmada com exame, que indica COVID caso qualquer um dos casos abaixo seja verdadeiro segundo os dados da tabela:

- Se o resultado da tomografia for ”típico COVID-19” , coluna ”TOMO_RES” = 1
- coluna ”RES_IGM” = 1, relacionada ao resultado do IGM
- colunas ”PCR_RESUL” = 1 e ”PCR_sars2” = 1, relacionadas ao PCR.

Veremos na análise de feature importance que este dado teve influência no resultado dos modelos de classificação.

A Fig. 8 mostra a relação entre a coluna criada e a coluna de classificação final (CLASSI_FIN) original da base.

A fig. 9 foi obtida a partir da divisão do número de casos internados na uti de cada doença pelo total de casos de cada doença. Esta porcentagem mostra o quanto cada doença piora tende a causar internação na UTI.

A ”taxa UTI ” (fig. 9) é diferente da ”ocupação UTI ” (fig. 10), uma vez que a ocupação indica quantos casos de UTI eram de cada doença; e a taxa indica, para cada doença, quantos foram para UTI. Uma doença pode ter maior porcentagem de ocupação de UTI do que outra apesar de menor tendência de necessitar de tratamento intensivo. Note que mais de 60% dos casos de internações em UTI foram de COVID-19 (Fig. 10). Porém, os casos classificados com COVID-19 não são aqueles que mais tenderam a ser internados na UTI (Fig. 9). Isso ocorre porque a quantidade de casos classificados com COVID-19 é muito alta (Fig. 7).

Os gráficos apresentados relacionados a esta seção contam os casos de todos os períodos que estavam na base (de 2020 até o início de 2022).

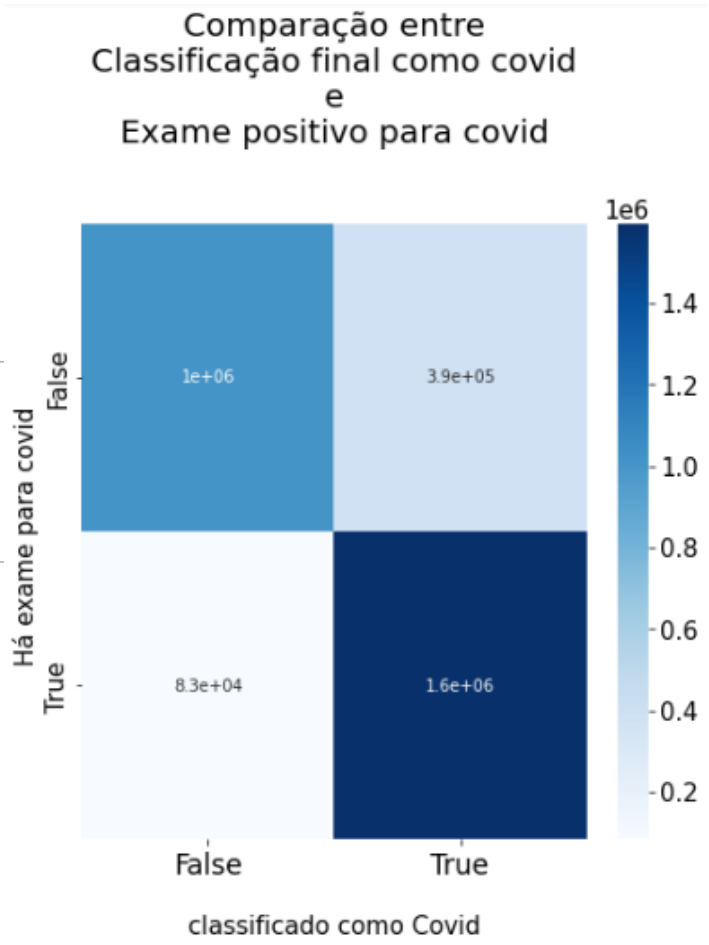

Fig. 8. Matriz de confusão dos casos de nossa coluna de confirmação por exame e da coluna de classificação final da base. (após limpeza da base)

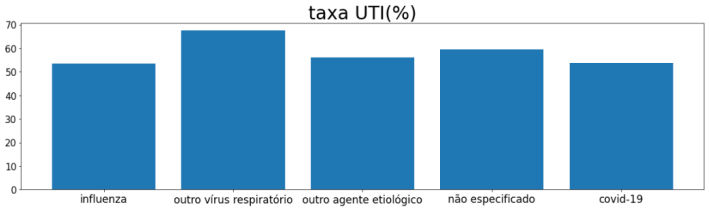

Fig. 9. porcentagem de pacientes que foram internados na UTI para cada (classificação final) doença (após limpeza da base).

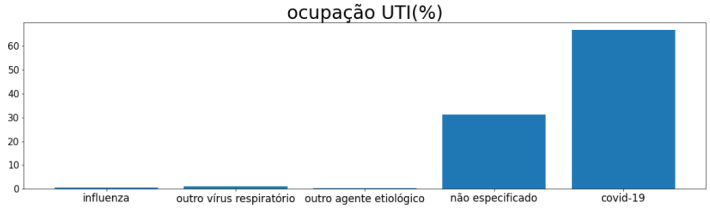

Fig. 10. quantidade (percentual) de cada caso na UTI(após limpeza da base)

D. 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 de pessoas por idade que receberam as doses da vacina Fig. 13.

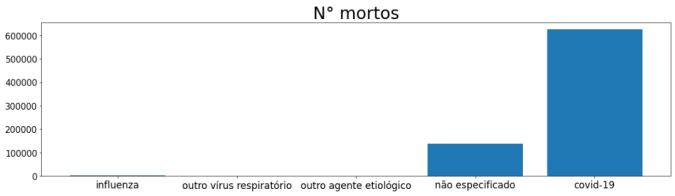

Fig. 11. quantidade de mortos para cada classificação final (após limpeza da base)

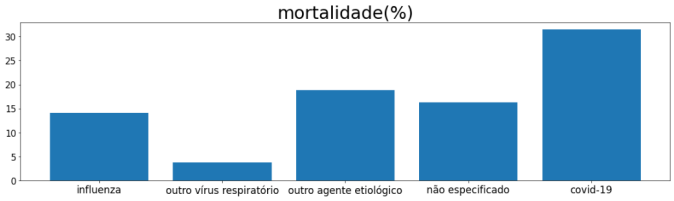

Fig. 12. quantidade de casos para cada classificação final (após limpeza da base)

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.

Para Fig. 13:

parâmetro	valor
Média de idade	65.02
desvio padrão da idade	18.27
faixa dentro de 1 desvio padrão	[46.75, 83.29]
faixa dentro de 2 desvios padrões	[28.47, 101.56]

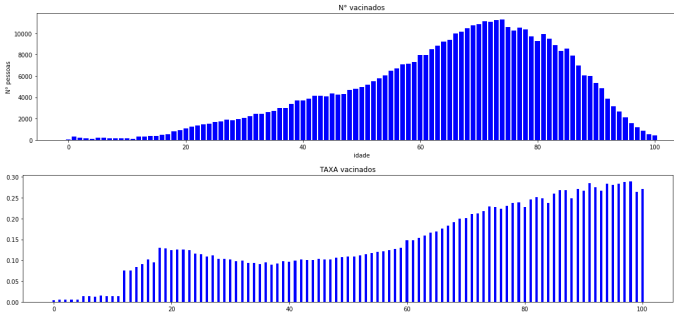

Fig. 13. Distribuição e taxa de vacinação por idade. As marcações no eixo y do gráfico de baixo (taxa de vacinação) vão de 0% até 30% (aumentando de 5 em 5).

2) eficácia e mortalidade: A Fig. 14 mostra a divisão do número de pessoas que tomaram a vacina e morreram de uma certa idade pelo número total de pessoas vacinadas desta idade. Além disso, foram considerados apenas os casos acima de 12 anos (para terem acesso a vacina) e a partir de uma data próxima a disponibilização da vacina (no caso 31/01/2021). Desse modo, os casos terão ocorrido no mesmo período de tempo, evitando que um grupo tenha sido exposto a uma variante que o outro não foi.

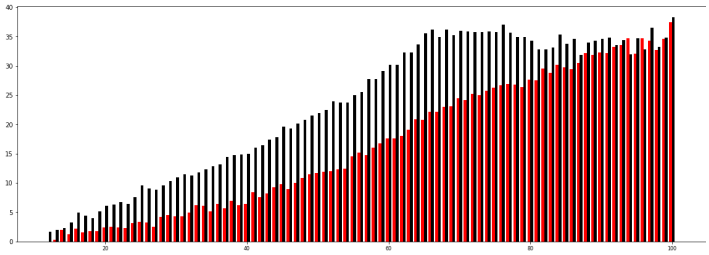

Fig. 14. Em vermelho: mortalidade de vacinados. Em preto: mortalidade de não vacinados. Os valores só consideram os casos de maiores de 12 anos e que ocorreram depois de 31/01/2021. As marcações de idade são: 20, 40 ,60, 80 e 100; e as de mortalidade são: 0,10,20,30 e 40 (em porcentagem)

Como as vacinas diminuem sua eficácia com o tempo [6], fizemos uma coluna que indica o tempo desde última dose tomada da vacina da COVID-19 até os primeiros sintomas: "TEMP_ULT_DOSE". Ela foi útil no treinamento de modelos de aprendizagem de máquina (que serão descritos em outras partes deste projeto).

3) gripe: Até aqui todos casos de vacina analisados foram relacionados à vacina da COVID-19.

Assim como a vacinação contra COVID-19 não estava registrada na maior parte dos casos, a minoria tinha vacina contra gripe registrada (Fig. 15).

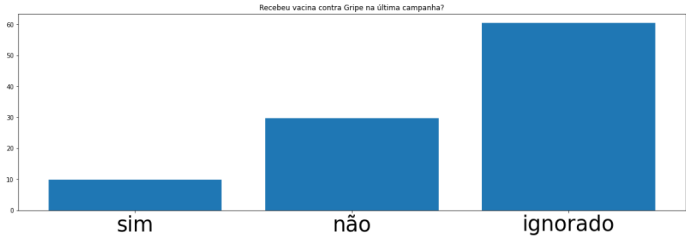

Fig. 15. Porcentagem de casos de acordo com a resposta ao questionamento: "Recebeu vacina contra Gripe na última campanha?" sim: 10%, não: 30%, ignorado: 60%

E. Inferência das variantes

As variantes de SARS-COV-2 que surgiram ao longo da pandemia ocasionaram os aumentos repentinos de casos observados na fig. 16 e na fig. 17, que são as distribuições diárias de casos (de SRAG e de SRAG por COVID-19) de acordo com a data dos primeiros sintomas.

A variante que predominava durante a data dos primeiros sintomas de cada paciente foi estimada visualmente a partir dos gráficos dos números de casos de COVID-19 por dia, que tiveram a imagem modificada para melhor visualização.

A distribuição do número de casos diários de SRAG classificados como COVID-19 (considerando o dia dos primeiros sintomas como o dia do caso) pode ser representada

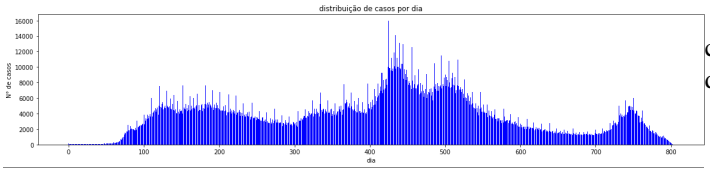

Fig. 16. N° de casos diários de SRAG (do dia 0 até o dia 804)

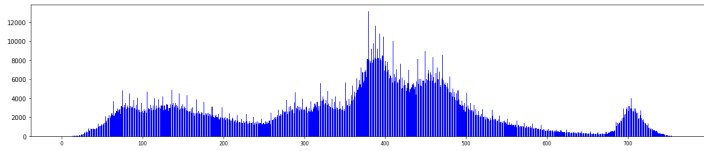

Fig. 17. N° de casos diários de SRAG classificados como COVID-19 (do dia 0 até o dia 804)

por uma função $f : [0, 804] \rightarrow [0, 14.000]$, em que $[0, 804]$ contém os dias possíveis na base, e o máximo de casos de COVID-19 em um dia foi inferior a 14.000.

A seguinte sequência de operações foi aplicada à função f:

1) operação linear de normalização (subtrair da média e dividir por uma constante) na função. A imagem da função normalizada passa a ser $[-1, 1]$. Chamemos esta função normalizada de f_{norm}.
2) Transformada de Fourier por meio do algoritmo Fast Fourier Transform (FFT) [2].
3) Filtragem dos maiores valores no *power spectrum* gerado como output da FFT. Ondas com valor inferior a certo limite são 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}.

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 fig. 18.

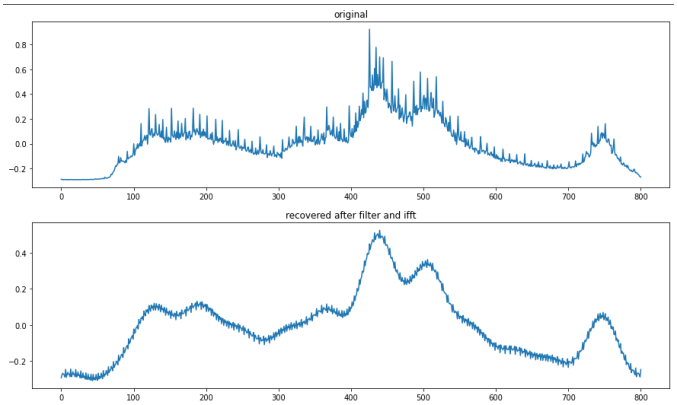

Fig. 18. f_{norm} (cima) e f_{filt} (baixo)

Um exemplo de onda pode ser visto na fig. 19 que corresponde a variante predominante no fim de 2021 e começo de 2022 (ômicron).

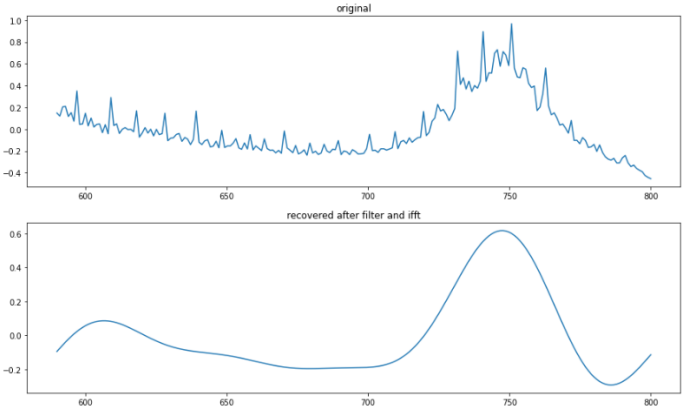

Fig. 19. f_{norm}(cima) e f_{filt}(baixo)

A partir da estimação visual do período de tempo de cada onda criamos a coluna ”ONDA” com cada linha possuindo um número que representa a(s) variante(s) predominante(s) no período que o paciente teve os primeiros sintomas. Esta coluna foi útil para os classificadores, como veremos em momento posterior neste artigo.

Por causa da falta de clareza do momento exato do começo de uma onda e fim da anterior, entre as datas 26/09/2020 e 07/11/2020 , consideramos que a variante era uma mistura das variantes B.1.1.33 e Zeta. Isso foi feito de modo que o número da onda fosse aumentando linearmente com o dia entre as datas. O número correspondente a variante começa com o número atribuído ao da variante da data 26/09/2020, que é o valor atribuído à B.1.1.33 (1), e terminando com o número atribuído à variante da data 07/11/2020, que é aquele que corresponde ao da variante Zeta (2). O mesmo foi feito para os intervalos entre [05/10/2020 , 30/11/2020] e [01/01/2020 , 07/03/2020].

Nº da Onda	Data Inicial	Data Final	Nome da Variante
(0,1)	01/01/2020	07/03/2020	sem variante
1	08/03/2020	25/09/2020	B.1.1.33
(1,2)	26/09/2020	07/11/2020	-
2	08/11/2020	02/02/2021	Zeta
3	03/02/2021	20/04/2021	Gama
4	21/04/2021	04/10/2021	Delta
(4,5)	05/10/2021	30/11/2021	-
5	01/12/2021	15/04/2022	Ômicron

Outras quantidades de ondas e datas correspondentes poderiam ser feitas, esta foi feita baseada em notícias ([1]).

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 testá-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-19 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 de nodes, 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-19 ou não para segunda rede). A resposta da rede para classificação é aquela cujo output do último layer possui maior valor. Diferentes tamanhos foram testados modelos (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 profundidade e número de neurônios, tentou-se melhorar sua qualidade com o modo de equilibrar os casos. Isto era necessário, porque a base, após ser limpa, possuía a grande maioria dos casos 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 as colunas mais genéricas, como: fator de risco , comorbidades, etc. E em outra situação usar as colunas específicas 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 não gerou mudança significativa, 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-19 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 a variadas técnicas de machine learning (ML), para isso utilizamos a biblioteca scikit-learning, que já possui os classificadores que iremos utilizar implementados. e também a biblioteca XGBoost, pois ela fornece uma estrutura interna capaz de aumentar os gradientes. Duas métricas diferentes para a acurácia foram usadas: accuracy score e cross value score.

1) Modelos de classificador: 5 classificadores diferentes foram testados:

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

2) 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%

3) Importância das features: Estimamos a importância das features por meio do regressor logístico e da 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 final fornecido pelo modelo. Como as variáveis foram normalizadas para uma mesma faixa de valores, esta premissa foi considerada verdadeira .

As 3 principais features para inferência da sobrevivência de acordo com todos os classificadores são, em ordem:

1) A classificação de COVID-19 confirmada por exame (CLASSI_COV_EXAME). Está feature não estava presente na base inicial, ela foi gerada da forma dita anteriormente neste artigo.
2) A idade do paciente, que estava na coluna NU_IDADE_N da base original.
3) A variante predominante da COVID-19 na época que o paciente teve os primeiros sintomas (coluna ONDA), que foi estimada pelo método descrito anteriormente .

Tabela de feature importance a partir de modelo de regressão logística

feature	coefficient
EXAME POSITIVO COVID	4.290000
IDADE	4.210000
ONDA (variante da COVID na época)	-1.800000
Tempo desde ultima dose contra COVID	1.510000
VACINA CONTRA COVID	1.100000
FATOR DE RISCO GRAVE	-0.780000
VACINA CONTRA GRIPE	-0.640000
FATOR DE RISCO	-0.550000
COMORBIDADE GRAVE	0.200000
UTI	0.140000

O modelo de árvore de decisão disponibilizado pelo sklearn fornece o atributo *feature_importances_* que foi usado para determinar a importância na tabela a seguir.

Note que a ordem das 4 primeiras features é a mesma que a anterior.

Tabela de feature importance a partir de modelo de árvore de decisão.

feature	importance
EXAME POSITIVO COVID	0.578
IDADE	0.2
ONDA (variante da COVID na época)	0.104
Tempo desde ultima dose contra COVID	0.035
VACINA CONTRA GRIPE	0.025
UTI	0.024
COMORBIDADE GRAVE	0.014
VACINA CONTRA COVID	0.012
FATOR DE RISCO	0.005
FATOR DE RISCO GRAVE	0.004

VI. Conclusão

Podemos recuperar, usando a rede neural com 90% de acurácia, os dados de classificação final de cerca de 33% da base de dados marcados com especificados (marcados como 4) ou não preenchidos corretamente (marcados com NaN). Usando a floresta randômica (97% de acurácia), podemos recuperar cerca de 14% da coluna EVOLUCAO que estava marcada como "Ignorado" (número 9 como valor).

References

[1] "avaliação das cinco ondas da covid-19 no rio de janeiro". In: *https://diariodepetropolis.com.br/integra/secretaria-mostra-estudo-com-avaliacao-das-cinco-ondas-da-covid-19-no-rj-207935*. 2022.

[2] 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.

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

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

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

[6] Wikipedia. "perda de eficácia". In: *https://www.cnnbrasil.com.br/saude/eficacia-das-vacinas-diminui-com-o-tempo-mas-ainda-oferecem-protecao-diz-estudo/*. 2022.