PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação Lato Sensu em Ciência de Dados e Big Data

DIAGNÓSTICO EM CASOS DE HOSPITALIZAÇÃO POR INFLUENZA (H3N2) OU COVID-19 ATRAVÉS DOS SINTOMAS

Cristiane Guimaraes Bastos Silva

O problema Proposto

A evolução da H3N2 no Brasil surpreendeu a população e as autoridades de saúde devido à ascensão repentina de casos de hospitalização durante o final de 2021 e início de 2022.

Esta ascensão da H3N2 se deu durante o período da pandemia de COVID-19, o que gerou muitas dúvidas sobre com qual vírus as pessoas estavam sendo contaminadas, pois os sintomas da variante Omicrôn, que surgiu em novembro de 2021, um mês antes da H3N2, tem sintomas muito semelhantes.

A motivação do tema deste trabalho foi a necessidade de ter um diagnóstico rápido para identificar qual a doença que o paciente estava contaminado,isto verificando os sintomas e estudando-os através de um algoritmo de machine learning para determinar sobre qual dessas 2 doenças a pessoa foi infectada.

Dados Utilizados

Foi utilizado o banco de dados, disponibilizado pelo Ministério da Saúde na plataforma do DataSUS, chamado de Banco de Dados de Síndrome Respiratória Aguda Grave (SRAG) - incluindo dados da COVID-19.

Este banco de dados faz o levantamento dos pacientes de hospitais e unidades de saúde que apresentam algum sintoma ou necessidade de hospitalização devido a uma síndrome respiratória.
(https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022)

Também foi analisada outra fonte de dados para integração no banco de dados, que é do Censo Demográfico de 2010 trazendo os valores de índice de Gini por município.
(https://censo2010.ibge.gov.br/sinopse/index.php?dados=7&uf=00)

Objetivos da Análise

Criar um algoritmo de classificação que consiga diagnosticar, com uma boa precisão, se pacientes hospitalizados por uma síndrome respiratória aguda grave foram infectados pelo vírus da Influenza (H3N2) ou coronavírus (COVID-19) através da análise dos sintomas manifestados.

Essa avaliação seria uma ótima alternativa para o paciente com a síndrome respiratória ter o diagnóstico o quanto antes.

Antecipando o diagnóstico clínico, que por vezes demora dias para se ter o resultado, pode-se iniciar o tratamento com medicação, pois, pode acontecer de perder um tempo precioso aguardando os resultados dos testes. Tempo esse importante, se a eficácia do algoritmo for comprovada, que pode salvar vidas.

Período de Análise e Limite Geográfico

Os dados analisados pertencem as 50 cidades que mais tiveram casos de hospitalização registrados por Síndrome Respiratória Aguda Grave (SRAG) no período de Dezembro de 2021 a Janeiro de 2022, época do pico da epidemia de H3N2 no Brasil.

Coleta de Dados

Importação das Bibliotecas necessárias

```
In [1]:
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
```

Coleta de dados e seleção de atributos da base do Ministério da Saúde

```
In [2]:
SRAG_22 = r'.....\datasets\INFLUD22-13-06-2022.csv'
dados_22dt = pd.read_csv(SRAG_22, delimiter=';',
usecols='DT_SIN_PRI CLASSI_FIN SEM_PRI FEBRE TOSSE GARGANTA DISPNEIA DESC_RESP SATURACAO DIARREIA VOMITO OUTRO_SIN DOR_ABD FADIGA PERD_OLFT PERD_PALA
encoding='ISO-8859-1')
dados_22dt.rename(columns={'CO_MUN_NOT':'IBGE'}, inplace=True)

dados_22 = dados_22dt

print(dados_22dt)
```

```
In [17]:
SRAG_21 = r'.....\datasets\INFLUD21-13-06-2022.csv'
dados_21dt = pd.read_csv(SRAG_21, delimiter=';',
usecols='DT_SIN_PRI CLASSI_FIN SEM_PRI FEBRE TOSSE GARGANTA DISPNEIA DESC_RESP SATURACAO DIARREIA VOMITO OUTRO_SIN DOR_ABD FADIGA PERD_OLFT PERD_PALA
encoding='ISO-8859-1')
dados_21dt.rename(columns={'CO_MUN_NOT':'IBGE'}, inplace=True)

dados_21 = dados_21dt

print(dados_21dt)
```

Coleta de dados e seleção de atributos da base do IBGE

```
In [65]:
file_gini = r'.....\datasets\datasets\gini.csv'
dados_gini = pd.read_csv(file_gini, delimiter=';',
encoding='ISO-8859-1')


print (dados_gini)
dados_gini.shape
dados_gini.head()
```

Processamento /Tratamento de Dados

	FEBRE	TOSSE	GARGANTA	DISPNEIA	DESC_RESP	SATURACAO	DIARREIA	VOMITO	OUTRO_SIN	DOR_ABD	FADIGA	PERD_OLFT	PERD_PALA
0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	1.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
100560	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
100561	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0
100562	0.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
100563	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
100564	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

100565 rows × 13 columns

<u>Base de Dados do Ministério da Saúde</u>

- Separação das colunas necessárias, no caso os sintomas.
- Check para identificar dados vazios e definido que tudo que era null passasse a ser 0.
- Retirada da base de dados todas as classificações de outras síndromes que não são de síndrome respiratória (1 e 5) e os registros de pacientes que não foram hospitalizados e resultado convertido em binário (0 e 1).
- Selecionado as 4 primeiras semanas (referentes as jan/22) e as 5 ultimas (referentes a dez/21).
- Junção das Bases de 2021 e 2022.
- Troca do nome da coluna “Codigo IBGE” para conseguir fazer a junção dos banco de dados do IBGE.
- Tranformação do dataset apenas em sintomas (possui sintoma =1, não possui sintoma=0) e indexação para ficar organizado.

Processamento /Tratamento de Dados

	Municipio	2010	IBGE
0	Alta Floresta D'Oeste	0,5893	110001
1	Alto Alegre dos Parecis	0,5491	110037
2	Alto Paraiso	0,5417	110040
3	Alvorada D'Oeste	0,5355	110034
4	Ariquemes	0,5496	110002
...	...	...	...
5560	Vianópolis	0,4672	522200
5561	Vicentinópolis	0,4824	522205
5562	Vila Boa	0,4935	522220
5563	Vila Propicio	0,524	522230
5564	Brasilia	0,637	530010

[5565 rows x 3 columns]

Base de Dados do IBGE

- Transformação dos campos texto concatenados para colunas.
- Alteração de virgula por ponto
- Definição de valores como float
- Seleção da média de GINI de 2010 e partir dela todos os valores que fossem menor que a média considerado 0 e os valores maiores que a média considerado 1.
- Verificação de quais cidades estavam acima e abaixo da média.

Processamento /Tratamento de Dados

Junção das bases de dados

Após a junção das bases de dados do SUS e IBGE, os valores de GINI foram transformados em binário (valores maiores que a média = 1 e menores que a média = 0.

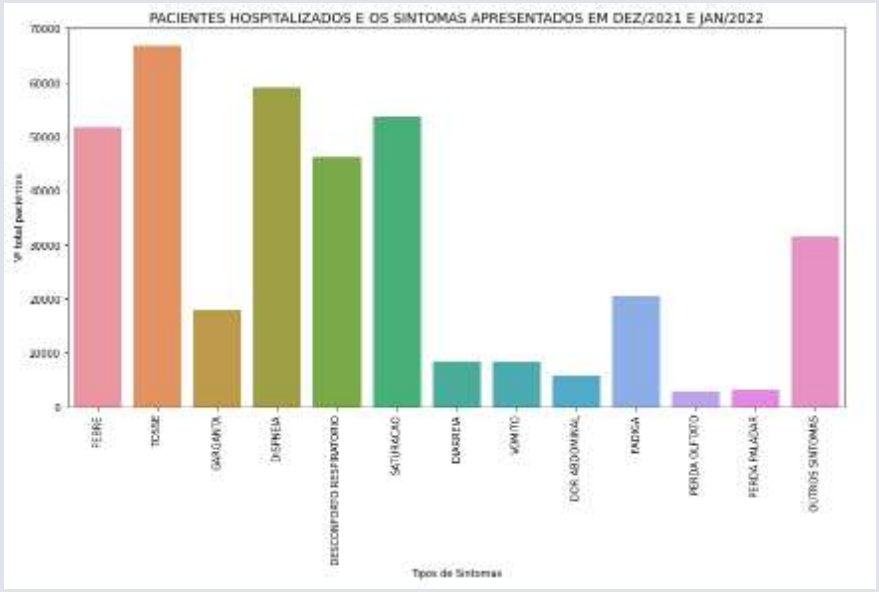
Separado como amostra as 50 cidades mais afetadas.

Salvo um novo dataset para utilização nos algoritmos de machine learnig

	FEBRE	TOSSE	GARGANTA	DISPNEIA	DESC_RESP	SATURACAO	DIARREIA	VOMITO	OUTRO_SIN	CLASSI_FIN	DOR_ABD	FADIGA	PERD_OLFT	PERD_PALA	2010
0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
1	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
2	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
3	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0
4	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
54831	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
54832	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
54833	0.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
54834	0.0	0.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
54835	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

54836 rows × 15 columns

Análise e Exploração dos Dados



Relação do número de pacientes hospitalizados e os sintomas apresentados em dezembro/21 e Janeiro/22.

Out[34]:

	Sintoma	Quantidade_Total	%
1	TOSSE	66678.0	66.30
3	DISPNEIA	59026.0	58.69
5	SATURACAO	53693.0	53.39
0	FEBRE	51663.0	51.37
4	DESCONFORTO RESPIRATORIO	46056.0	45.80
12	OUTROS SINTOMAS	31458.0	31.28
9	FADIGA	20503.0	20.39
2	GARGANTA	17831.0	17.73
6	DIARREIA	8417.0	8.37
7	VOMITO	8165.0	8.12
8	DOR ABDOMINAL	5717.0	5.68
11	PERDA PALADAR	3142.0	3.12
10	PERDA OLFTATO	2947.0	2.93

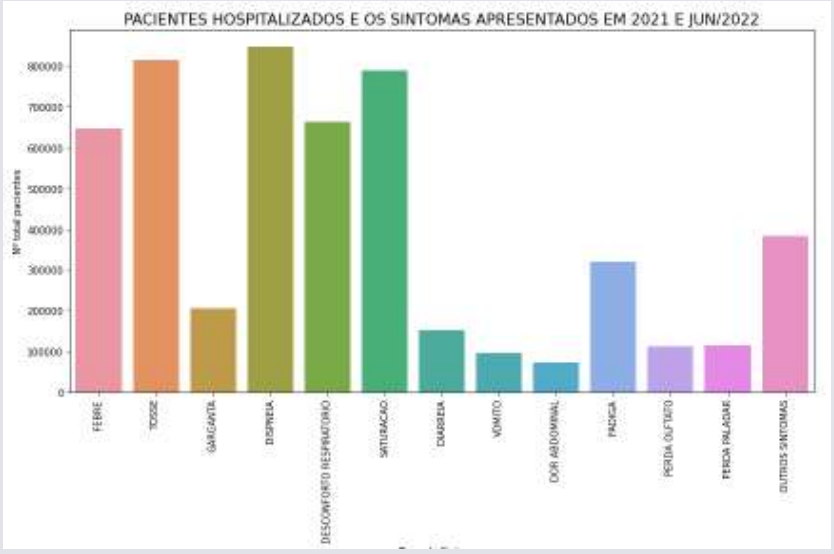
Out[37]:

	DT_SIN_PRI	SEM_PRI	IBGE	FEBRE	TOSSE	GARGANTA	DISPNEIA	DESC_RESP	SATURACAO	DIARREIA	VOMITO	OUTRO_SIN	HOSPITAL	CLASSI_FIN	DOR_ABD	FAD
0	04/01/2021	1	270430	1.0	1.0	2.0	1.0	2.0	1.0	2.0	2.0	2.0	1.0	5.0	2.0	
1	03/01/2021	1	500270	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	1.0	5.0	2.0	
2	03/01/2021	1	250750	1.0	1.0	NaN	1.0	NaN	NaN	NaN	NaN	1.0	1.0	4.0	NaN	
3	08/01/2021	1	410480	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0	1.0	4.0	2.0	
4	05/01/2021	1	351880	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0	1.0	5.0	2.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
303359	23/05/2022	21	310620	2.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	1.0	4.0	2.0	
303360	31/05/2022	22	412350	2.0	2.0	2.0	1.0	1.0	1.0	2.0	2.0	NaN	1.0	NaN	2.0	
303361	30/05/2022	22	355030	2.0	2.0	2.0	2.0	1.0	1.0	2.0	2.0	2.0	2.0	4.0	2.0	
303362	06/06/2022	23	355030	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	1.0	1.0	5.0	NaN	
303363	11/06/2022	23	354980	2.0	1.0	2.0	1.0	1.0	1.0	2.0	2.0	2.0	1.0	NaN	2.0	

2032195 rows × 18 columns

Análise e Exploração dos Dados

Relação do número de pacientes hospitalizados e os sintomas apresentados em 2021 a Junho/22



Out[40]:

	Sintoma	Quantidade_Total	%
3	DISPNEIA	847080.0	81.06
1	TOSSE	815791.0	78.78
5	SATURACAO	788927.0	78.11
4	DESCONFORTO RESPIRATORIO	662807.0	69.13
0	FEBRE	646626.0	65.49
12	OUTROS SINTOMAS	382972.0	45.92
9	FADIGA	320051.0	37.91
2	GARGANTA	205498.0	24.74
6	DIARREIA	150881.0	18.49
11	PERDA PALADAR	114139.0	14.26
10	PERDA OLFTATO	111831.0	13.97
7	VOMITO	96319.0	12.04
8	DOR ABDOMINAL	72225.0	9.17

In [41]:
```
print(merged_df)
```

```
        DT_SIN_PRI  SEM_PRI    IBGE  FEBRE  TOSSE  GARGANTA  DISPNEIA  \
1       03/01/2021        1  500270    1.0    1.0       1.0       1.0
4       05/01/2021        1  351880    2.0    2.0       2.0       1.0
9       05/01/2021        1  330455    NaN    1.0       NaN       1.0
14      03/01/2021        1  351280    1.0    1.0       1.0       1.0
18      03/01/2021        1  355030    1.0    1.0       2.0       2.0
...            ...      ...     ...    ...    ...       ...       ...
303338  04/05/2022       18  354880    1.0    1.0       NaN       1.0
303352  30/05/2022       22  410480    2.0    1.0       2.0       1.0
303353  04/06/2022       22  352590    1.0    NaN       NaN       NaN
303356  02/06/2022       22  431440    2.0    1.0       1.0       1.0
303362  06/06/2022       23  355030    NaN    NaN       NaN       NaN

        DESC_RESP  SATURACAO  DIARREIA  VOMITO  OUTRO_SIN  HOSPITAL  \
1             1.0        1.0       1.0     1.0        2.0       1.0
4             1.0        1.0       2.0     2.0        2.0       1.0
9             1.0        1.0       NaN     NaN        NaN       1.0
14            1.0        1.0       2.0     1.0        NaN       1.0
18            2.0        1.0       NaN     NaN        2.0       1.0
...           ...        ...       ...     ...        ...       ...
303338        NaN        1.0       NaN     NaN        NaN       1.0
303352        1.0        2.0       2.0     1.0        2.0       1.0
303353        NaN        NaN       NaN     NaN        NaN       1.0
303356        1.0        2.0       2.0     2.0        2.0       1.0
303362        NaN        NaN       1.0     NaN        1.0       1.0

        CLASSI_FIN  DOR_ABD  FADIGA  PERD_OLFT  PERD_PALA
1              1.0      2.0     1.0        1.0        1.0
4              1.0      2.0     2.0        2.0        2.0
9              1.0      NaN     1.0        NaN        NaN
14             1.0      2.0     1.0        1.0        1.0
18             1.0      NaN     1.0        2.0        2.0
...            ...      ...     ...        ...        ...
303338         1.0      NaN     NaN        NaN        NaN
303352         1.0      2.0     2.0        2.0        2.0
303353         1.0      1.0     NaN        NaN        NaN
303356         1.0      2.0     2.0        2.0        2.0
303362         1.0      NaN     NaN        NaN        NaN

[1184730 rows x 18 columns]
```

Análise e Exploração dos Dados

Percebe-se que numero de casos com sintomas de 2021 a junho/22 são diferentes dos do período de dezembro/21 a janeiro/22 o que indica alguma anomalia no período, significando que a hipótese apresentada é real e há necessidade de estudo.

Out[34]:

	Sintoma	Quantidade_Total	%
1	TOSSE	66678.0	66.30
3	DISPNEIA	59026.0	58.69
5	SATURACAO	53693.0	53.39
0	FEBRE	51663.0	51.37
4	DESCONFORTO RESPIRATORIO	46056.0	45.80
12	OUTROS SINTOMAS	31458.0	31.28
9	FADIGA	20503.0	20.39
2	GARGANTA	17831.0	17.73
6	DIARREIA	8417.0	8.37
7	VOMITO	8165.0	8.12
8	DOR ABDOMINAL	5717.0	5.68
11	PERDA PALADAR	3142.0	3.12
10	PERDA OLFTATO	2947.0	2.93

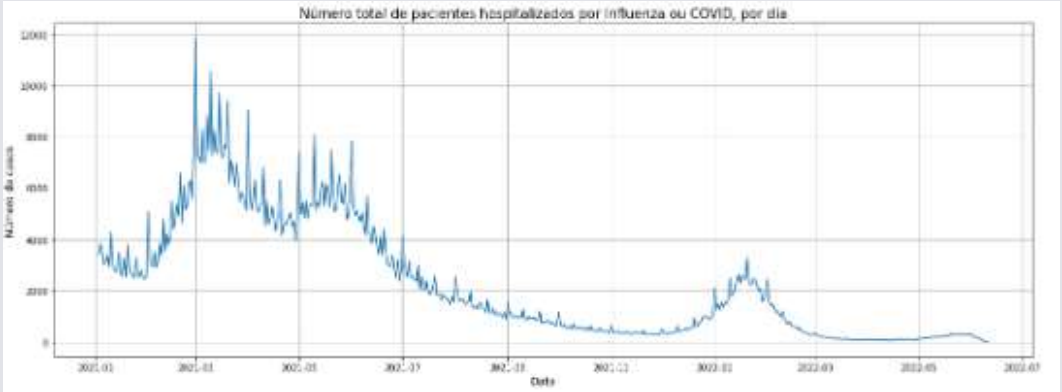
Out[40]:

	Sintoma	Quantidade_Total	%
3	DISPNEIA	847080.0	81.06
1	TOSSE	815791.0	78.78
5	SATURACAO	788927.0	78.11
4	DESCONFORTO RESPIRATORIO	662807.0	69.13
0	FEBRE	646626.0	65.49
12	OUTROS SINTOMAS	382972.0	45.92
9	FADIGA	320051.0	37.91
2	GARGANTA	205498.0	24.74
6	DIARREIA	150881.0	18.49
11	PERDA PALADAR	114139.0	14.26
10	PERDA OLFTATO	111831.0	13.97
7	VOMITO	96319.0	12.04
8	DOR ABDOMINAL	72225.0	9.17

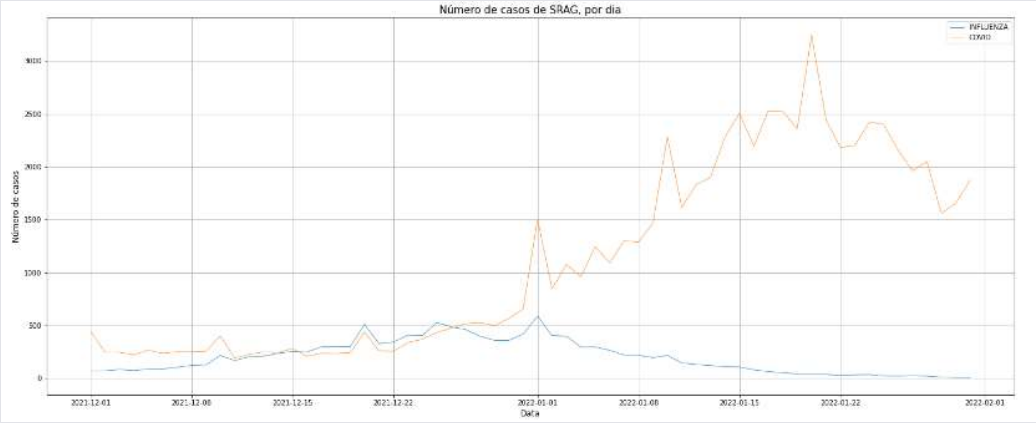
Análise e Exploração dos Dados

Para profundar na análise dos dados fiz os gráficos de todos os pacientes hospitalizados com Covid19 e H3N2 por dia.

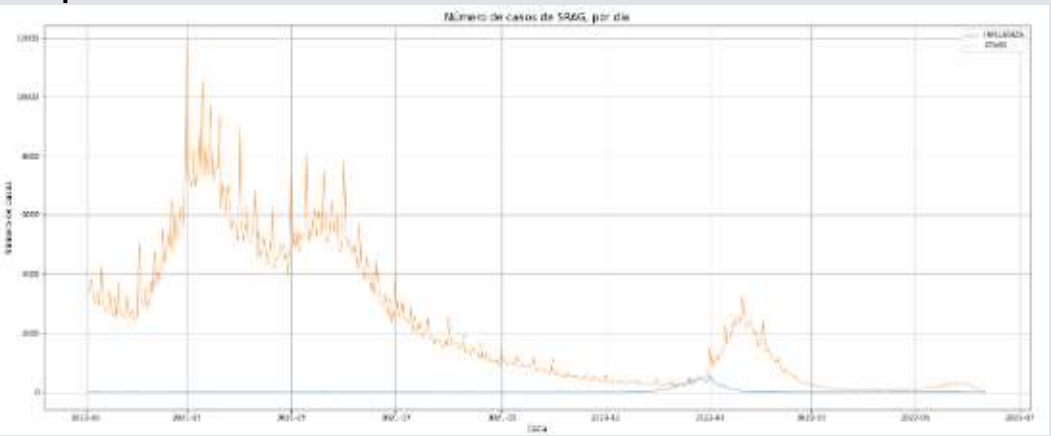
Total de casos de hospitalização somados de H3N2 e Covid de 2021 a junho/22



Total de casos de hospitalização de H3N2 e Covid dez/21 a jan/22

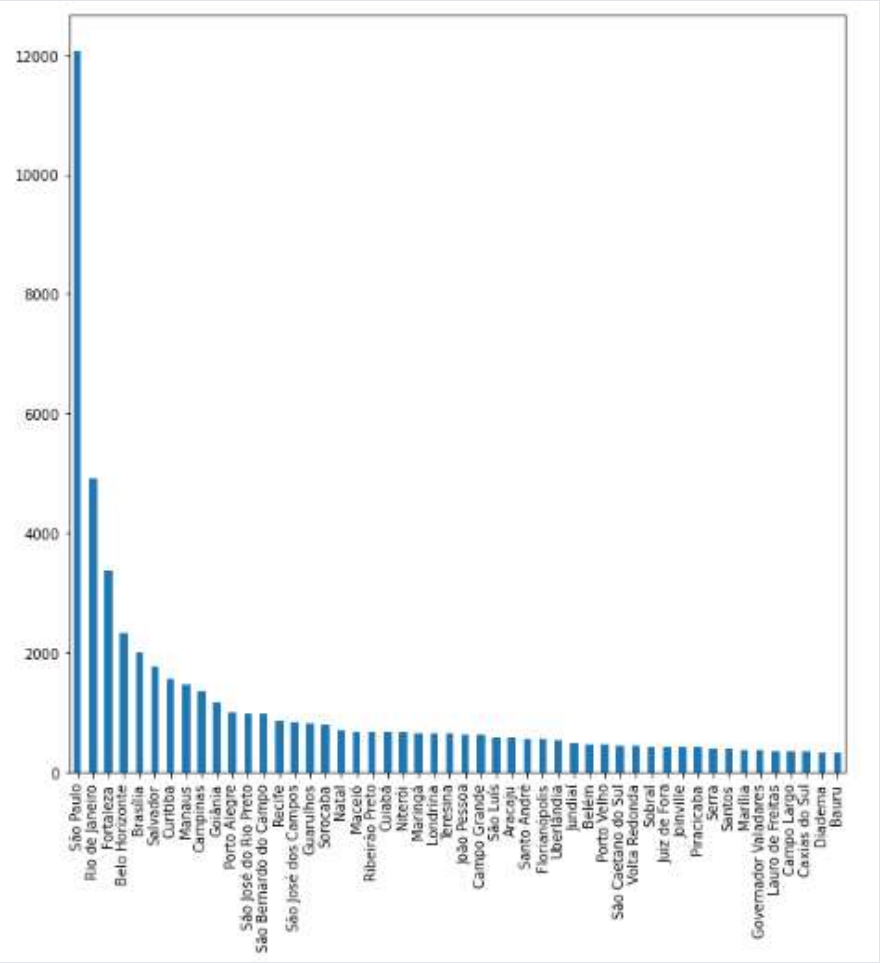


Total de casos de hospitalização de H3N2 e Covid de 2021 a junho/22 separadamente.



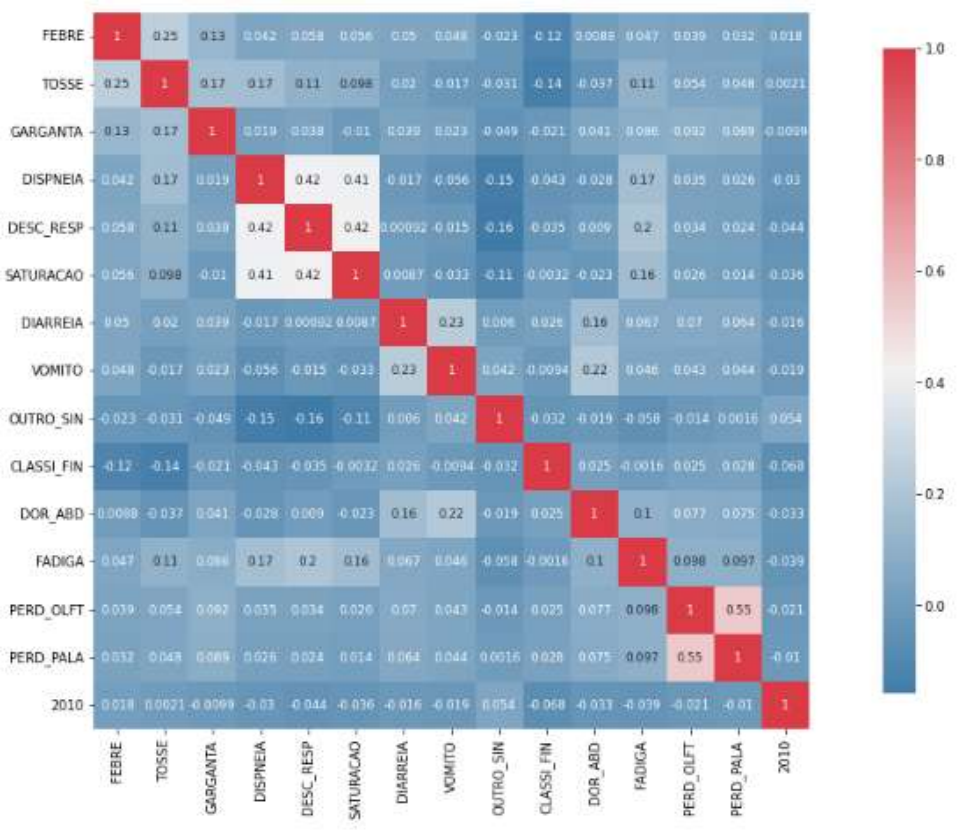
Análise e Exploração dos Dados

Casos de hospitalização por Covid e H3N2 nas 50 cidades mais afetadas



Análise e Exploração dos Dados

Com o mapa de correlação percebe-se que a perda do olfato e paladar tem uma forte correlação. Outro ponto é dispneia com descrição respiratória, saturação. Diante disso já conseguimos identificar alguns padrões que as doenças apresentam.



Criação de Modelos de Machine Learnig

Após o tratamento e análise dos dados seguimos para os algoritmos de machine learnig

Primeiramente foi feita a importação das Bibliotecas necessárias e do banco de dados criado na etapa de tratamento de dados.

Importação das bibliotecas

```
In [1]:
import pandas as pd
import numpy as np
import imblearn
from sklearn import metrics
from scipy import stats
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import shapiro
from matplotlib import rc
%matplotlib inline
from imblearn.under_sampling import NearMiss
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import accuracy_score, f1_score, recall_score, precision_score, confusion_matrix, classification_report
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn import model_selection
from sklearn.model_selection import cross_val_score, KFold, train_test_split, GridSearchCV
from sklearn.metrics import roc_auc_score, roc_curve, classification_report, accuracy_score, confusion_matrix
from yellowbrick.classifier import ClassificationReport
from yellowbrick.classifier import ROCAUC
from yellowbrick.classifier import ClassPredictionError
import scikitplot as skplt
import warnings
warnings.filterwarnings('ignore')
```

Importação do banco de dados preparado

```
In [2]:
SAMPLE_SET = r'.....\datasets\SAMPLE_SET.csv'
SAMPLE_SET = pd.read_csv(SAMPLE_SET, index_col=0)
SAMPLE_SET.rename(columns={'CLASSI_FIN':'target'}, inplace=True)
print (SAMPLE_SET)
SAMPLE_SET.shape
```

Criação de Modelos de Machine Learnig

Definido o target, foi feita a separação entre o modelo de treino e o modelo de teste.

```
In [4]:
X = df.drop("target", axis = 1)
y = df.target
xshape, yshape = df.shape
print("O dataframe possui {} amostras (linhas) e {} variáveis (colunas)".format(xshape,yshape))

O dataframe possui 54836 amostras (linhas) e 15 variáveis (colunas)
```

```
In [5]:
x_treino, x_teste, y_treino, y_teste = train_test_split(X, y, test_size = 0.25, random_state=5)
```

```
In [6]:
x_treino1 = x_treino
x_teste1 = x_teste
y_treino1 = y_treino
y_teste1 = y_teste
```

Amostras de treino e teste

```
x_treino:  (41127, 14)
x_teste:  (13709, 14)
y_treino:  (41127,)
y_teste:  (13709,)
```

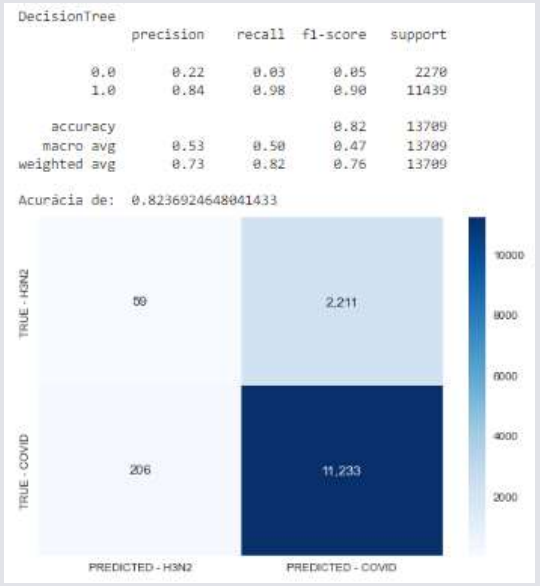
Classificações com Árvore de decisão, Random Forest, KNN (K-Nearet Neighbors e Regressão Logica.

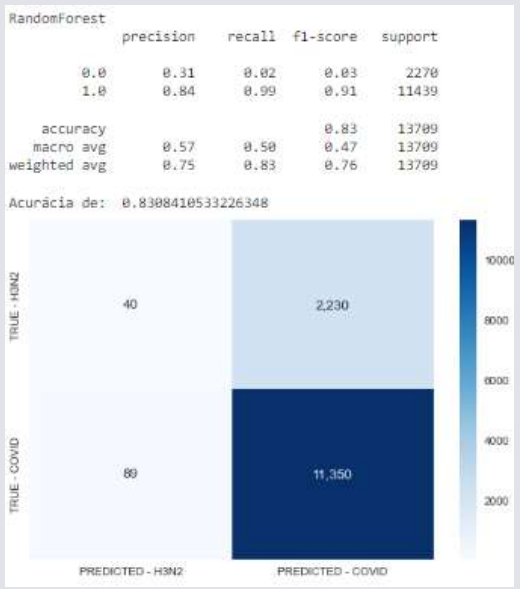
```
In [9]:
classificadores = {
    "DecisionTree": DecisionTreeClassifier(random_state=5),
    "RandomForest": RandomForestClassifier(random_state=5),
    "KNN": KNeighborsClassifier(),
    "LogisticRegression": LogisticRegression(random_state=5)}
```

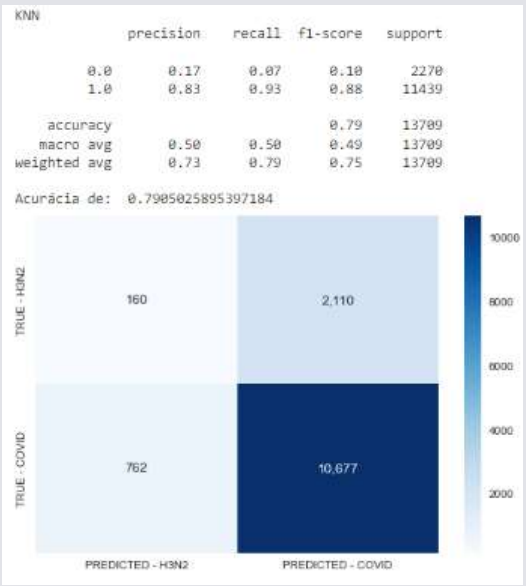
```
In [10]:
for nome_modelo in classificadores:
    modelo = classificadores[nome_modelo]
    modelo.fit(x_treino, y_treino)
    previsoes = modelo.predict(x_teste)
    avalia_metricas(y_teste, previsoes, nome_modelo)
    classificadores[nome_modelo] = modelo
```

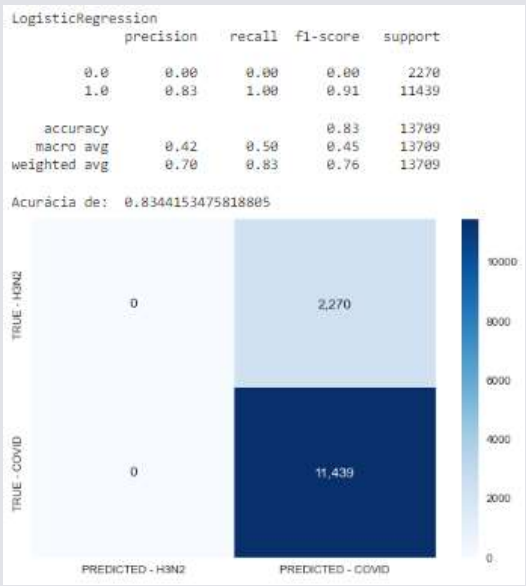
Criação de Modelos de Machine Learnig

Como 90% do banco de dados é de Covid, o algoritmo identificou muitas amostras da classe COVID e pouca da classe H3N2 com uma acurácia de 83% o que demonstra que a classe majoritária (Covid) teve maior influencia no algorítimo diante da classe minoritária. Com isso o algoritmo vai tender a achar que todos os casos são Covid.



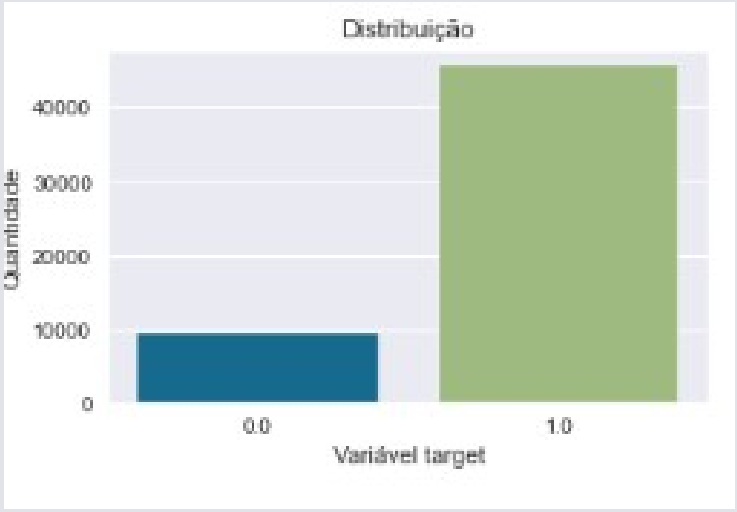




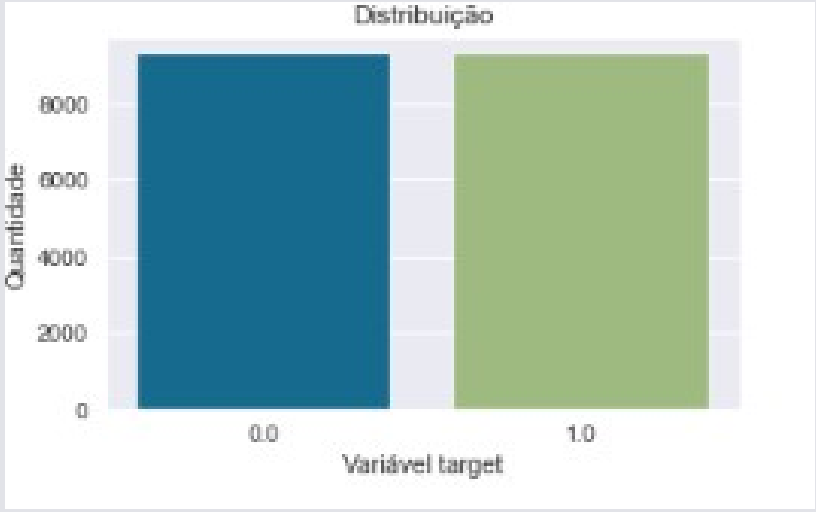


Criação de Modelos de Machine Learnig

Algoritmos identificou muitas amostras da classe COVID (1) classe majoritária e pouca da classe H3N2 (0), classe minoritária.

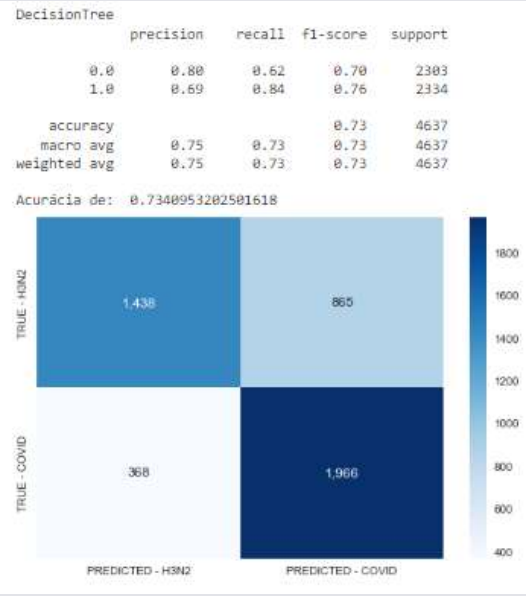


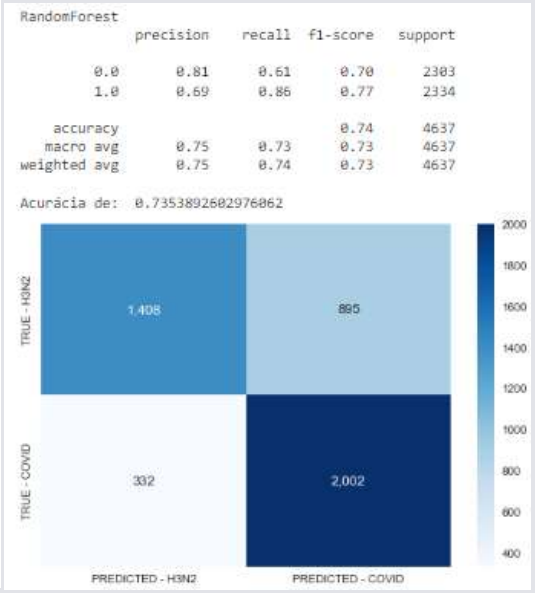
Para ter uma correlação correta foi necessário fazer o balanceamento das classes, selecionando aleatoriamente a mesma quantidade de amostras da classe majoritária para ficar igual ao numero de amostras da classe minoritária.

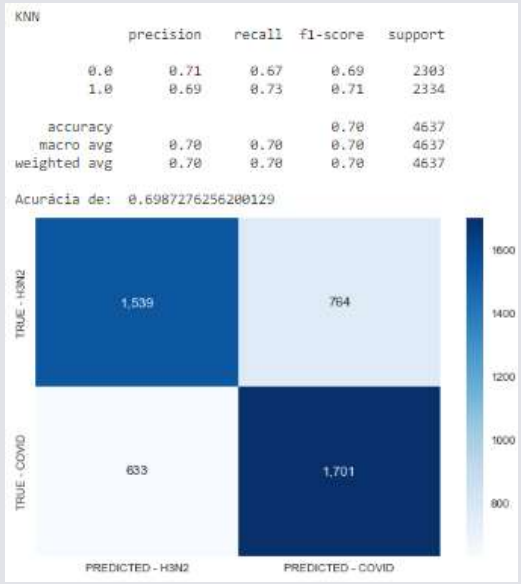


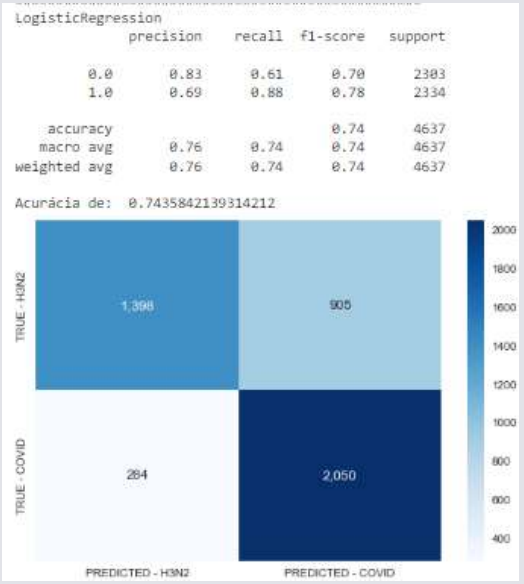
Criação de Modelos de Machine Learnig

Repetiu-se todo o processo de classificação após o balanceamento das classes.
Percebe-se que há uma boa acurácia de em média 73%.









Criação de Modelos de Machine Learnig

Através do model tuning foi obtido os parâmetros dos modelos que possuem os melhores resultados para trazer robustez ao algoritmo.
Foi utilizado o método de validação cruzada, kfold fazendo-se o cálculo da acurácia do modelo gerando uma média de resultados.

Média de desempenho dos modelos

```
Arvore           0.736642
Random forest    0.742180
KNN              0.729649
Logistica        0.738378
dtype: float64
```

In [29]:

```
resultados_arvore = []
resultados_random_forest = []
resultados_knn = []
resultados_logistica = []
for i in range(30):
    print(i)
    kfold = KFold(n_splits=10, shuffle=True, random_state=i)

    arvore = DecisionTreeClassifier(criterion='gini', min_samples_leaf=10, min_samples_split=2, splitter='random')
    scores = cross_val_score(arvore, X_df, Y_df, cv = kfold)
    resultados_arvore.append(scores.mean())

    random_forest = RandomForestClassifier(criterion = 'entropy', min_samples_leaf=10, min_samples_split=5, n_estimators=40)
    scores = cross_val_score(random_forest, X_df, Y_df, cv = kfold)
    resultados_random_forest.append(scores.mean())

    knn = KNeighborsClassifier(algorithm='brute', n_neighbors=20, p=1, weights='uniform')
    scores = cross_val_score(knn, X_df, Y_df, cv = kfold)
    resultados_knn.append(scores.mean())

    logistica = LogisticRegression(C = 1.0, multi_class = 'multinomial', solver = 'sag', tol = 0.0001)
    scores = cross_val_score(logistica, X_df, Y_df, cv = kfold)
    resultados_logistica.append(scores.mean())
```

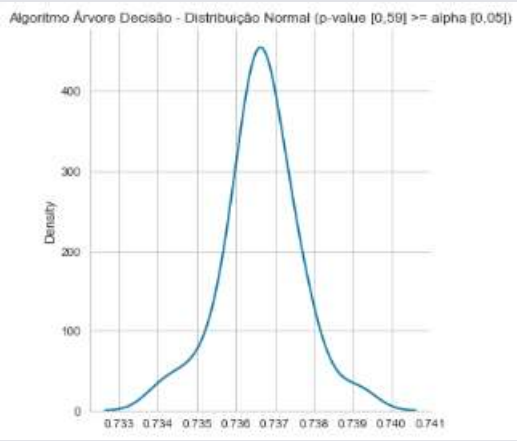
	Arvore	Random forest	KNN	Logistica
count	30.000000	30.000000	30.000000	30.000000
mean	0.736642	0.742180	0.729649	0.738378
std	0.000973	0.000798	0.001387	0.000372
min	0.734120	0.740322	0.726141	0.737464
25%	0.736290	0.741885	0.729052	0.738178
50%	0.736628	0.742100	0.729564	0.738353
75%	0.737046	0.742707	0.730711	0.738637
max	0.739134	0.743718	0.731749	0.739242

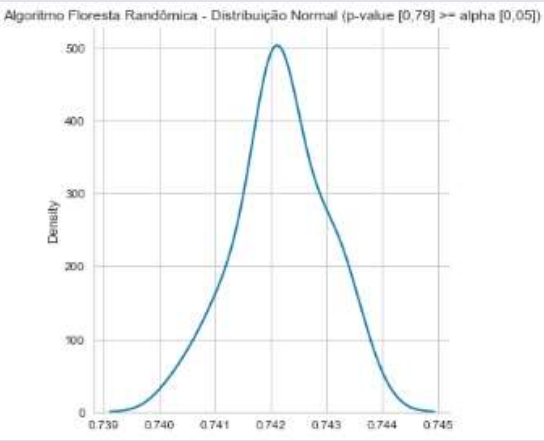
Interpretação dos Resultados

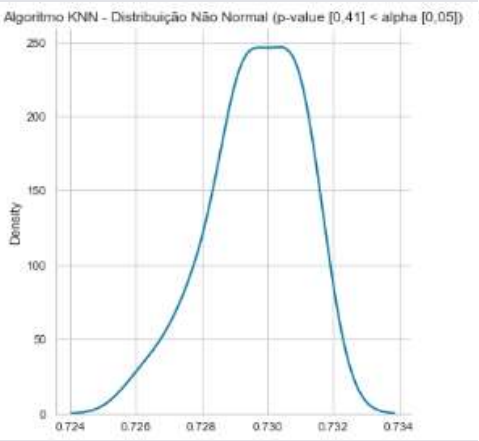
Os testes de normalidade tem como objetivo avaliar se uma distribuição de um conjunto de dados de uma variável aleatória é semelhante a uma distribuição normal.

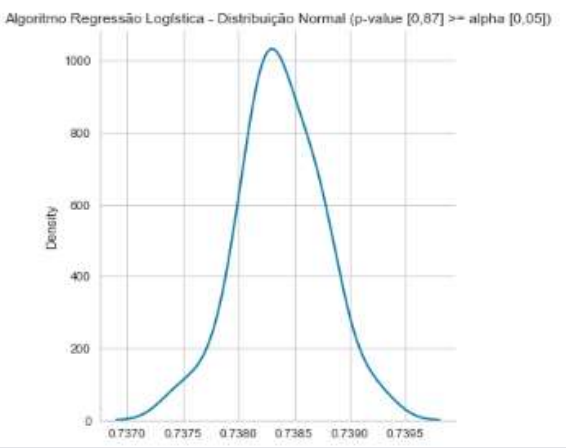
Para o teste de normalidade utilizamos o Teste de Shapiro-Wilk.

Os resultados dos testes de normalidade aplicados apontaram uma distribuição normal.



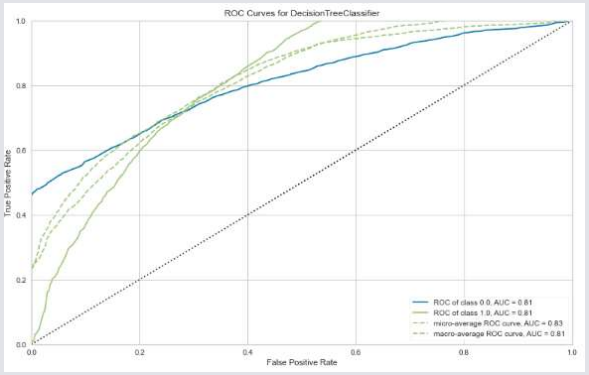


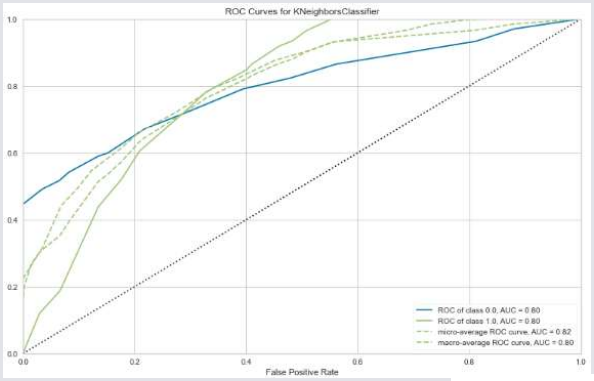


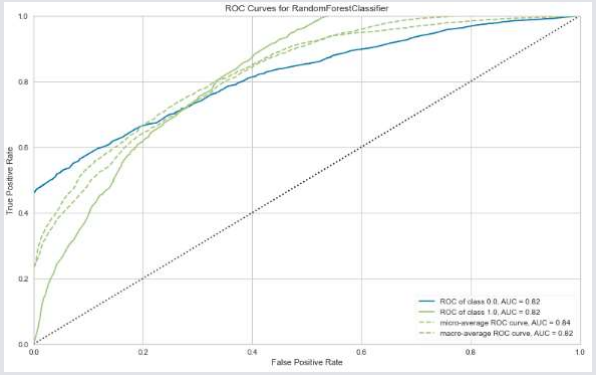


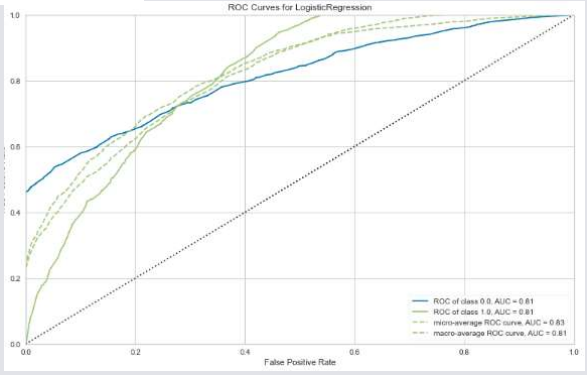
Interpretação dos Resultados

Parametriza-se novamente pelo GridSearchCV os modelos de machine learning com os valores ótimos para avaliação da curva ROC.
A curva ROC é uma medida de desempenho para verificar o quanto o modelo é capaz de distinguir entre as classes.
Quanto mais acentuada é a curva, melhor será o modelo em distinguir entre pacientes hospitalizados com COVID-19 ou H3N2









Conclusão

Com este algoritmo é possível trazer um 'diagnóstico artificial', a partir da análise dos sintomas registrados nos casos em que o paciente foi hospitalizado, com uma acurácia de cerca de 70%, diminuindo a dúvida se a pessoa estava contaminada por Covid19 ou H3N2.

Links

Link para o vídeo:

- https://www.youtube.com/watch?v=kqsEGu_nx5k

Link para o repositório

- https://github.com/cristiane-silva/TCC_PUC_MINAS

Link para os datasets:

- https://opendatasus.saude.gov.br/dataset/srag-2021-e-2022
- https://censo2010.ibge.gov.br/sinopse/index.php?dados=7&uf=00)