PREVER ESTÁGIO DE COVID19

Nury Yuleny Arosquipa Yanque
NUSP: 9871836

Resumo

Utilizei o dataset disponibilizado pelo GRUPO FLEURY, tem basicamente dois arquivos, um contendo os pacientes e dados associados ao paciente, e o outro arquivo contém basicamente os resultados dos exames por cada paciente e data com que foi feita. A partir disso, processou-se os dados com a finalidade de alimentar uma rede neural profunda por cada estágio da COVID19 (PCR, IgG e IgM), no total 3 redes. Foi muito legal conhecer a adversidade com que se lida quando se trabalha com dados do mundo real, gastei um tempão nisso, não consegui colocar nas CNNs.

Introdução

Pessoalmente eu tive duas motivações principais para resolver este EP. Primeiro, conseguir lidar com dataset do mundo real, quer dizer, entender os dados e arrumar ela de forma que possa ser processada por uma rede neural. Segundo, a partir dos dados que podem ser vários exames e dados do paciente, utilizar uma rede neural (na verdade 3 redes neurais, uma por cada estágio do COVID) conseguir prever qual o estágio do COVID19 num paciente.

Os objetivos do trabalho, para mim, são:

- Processar os dados, isto é, quais são os rótulos? quais que vão ser minhas features? isso conlleva a limpar os dados, transformar os dados, estruturá-los também, etc.
- Criar uma rede neural convolucional por cada estágio do Covid, isto é:
    - CNN para PCR
    - CNN para detecção de anticorpos IgG
    - CNN para detecção de anticorpos IgM

Vamos manter a estrutura proposta pelo EP4.

Metodologia

Pré-processamento dos dados

Estoy trabalhando com os dados disponibilizados pelo grupo Fleury, que são:

- Grupo_Fleury_Dataset_Covid19_Resultados_Exames.csv
- Grupo_Fleury_Dataset_Covid19_Pacientes.csv

Vamos analisar os dois arquivos

Grupo_Fleury_Dataset_Covid19_Pacientes

Tem 129596 registros com 7 colunas:

```
pacientes.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 129596 entries, 0 to 129595
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   ID_PACIENTE     129596 non-null  object
 1   IC_SEXO         129596 non-null  object
 2   AA_NASCIMENTO   129596 non-null  object
 3   CD_PAIS         129596 non-null  object
 4   CD_UF           129596 non-null  object
 5   CD_MUNICIPIO    129596 non-null  object
 6   CD_CEP          129596 non-null  object
dtypes: object(7)
memory usage: 7.9+ MB
```

- Não tem registros duplicados
- Adicionamos uma nova coluna ‘Idade’, que calcula o ano do paciente utilizando o ‘AA_NASCIMENTO’. Quando o valor do ano é ‘AAAA’ coloca 95 anos como idade, pois ‘AAAA’ é o valor asignado quando o ano de nascimento é igual ou anterior a 1930.
- Passou dados categoricos para numericos como os da coluna ‘IC_SEXO’ e ‘CD_UF’.
- Deletou-se as seguintes colunas: ['AA_NASCIMENTO', 'CD_PAIS', 'CD_UF', 'CD_MUNICIPIO', 'CD_CEP'].
- Todas estas mudanças foram salvas no arquivo: **Covid19_Pacientes_clean.csv**

Grupo_Fleury_Dataset_Covid19_Resultados_Exames

- Tem 2 496 591 registros com 8 colunas
- Foram deletados os registros duplicados
- Obteve-se a densidade dos exames, isto é, tendo em conta as colunas [‘DE_EXAME’, ‘DE_ANALITO’] a quantidade de cada DE_ANALITO. O resultado foi salvo no arquivo **examens_agrupados.csv**
- A partir disso, se deleteram os seguintes DE_ANALITOS:
    - 'COVID 19, Material'

  E os IgM e IgG não relacionados com a COVID19
    - 'CITOMEGALOVIRUS, ANTICORPOS IgG, soro' 'CMV, IgG'
    - 'CITOMEGALOVIRUS, ANTICORPOS IgM, soro' 'CMV, IgM'
    - 'RUBEOLA, ANTICORPOS IgG, soro' 'Rubeola, IgG'
    - 'RUBEOLA, ANTICORPOS IgM, soro' 'Rubeola, IgM'
    - 'VIRUS EPSTEIN BARR (ANTIGENO DO CAPSIDEO VIRAL), ANTICORPOS IgM, soro' 'Epstein-Barr, IgM'
    - 'VIRUS EPSTEIN BARR (ANTIGENO DO CAPSIDEO VIRAL), ANTICORPOS IgG, soro' 'Epstein-Barr, IgG'

- 'ZIKA VÍRUS, PESQUISA DE ANTICORPOS DA CLASSE IGM, soro' 'Zika vírus - IgM'
- 'ZIKA VÍRUS, PESQUISA DE ANTICORPOS DA CLASSE IGM, soro' 'Zika vírus - IgM - Índice'
- 'HERPES SIMPLEX, TIPO 1, ANTICORPOS IgG, soro' 'Herpes Simplex, Tipo 1, IgG'
- 'HERPES SIMPLEX, TIPO 2, ANTICORPOS IgG, soro' 'Herpes Simplex, Tipo 2, IgG'- 'SARAMPO, ANTICORPOS IgG, soro' 'Sarampo, Anticorpos IgG'
- 'SARAMPO, ANTICORPOS IgM, soro' 'Sarampo, Anticorpos IgM'
- 'CHLAMYDIA TRACHOMATIS, ANTICORPOS IgG, soro' 'Chlamydia Trachomatis, IgG'
- 'CHLAMYDIA TRACHOMATIS, ANTICORPOS IgM, soro' 'Chlamydia Trachomatis, IgM'
- 'VARICELLA-ZOSTER VIRUS, ANTICORPOS IgG, soro' 'Varicella-Zoster, anticorpos IgG'
- 'VARICELLA-ZOSTER VIRUS, ANTICORPOS IgM, soro' 'Varicella-Zoster, anticorpos IgM'
- 'ZIKA VÍRUS, PESQUISA DE ANTICORPOS DA CLASSE IGG, soro' 'Zika vírus - IgG - Índice'
- 'ZIKA VÍRUS, PESQUISA DE ANTICORPOS DA CLASSE IGG, soro' 'Zika vírus - IgG'
- 'DENGUE, ANTICORPOS IgG, PROVA RÁPIDA, soro' 'Dengue, IgG'
- 'DENGUE, ANTICORPOS IgM, PROVA RÁPIDA, soro' 'Dengue, IgM
- 'DENGUE, ANTICORPOS IgG, soro' 'Dengue, IgG'
- 'DENGUE, ANTICORPOS IgM, soro' 'Dengue, IgM'
- 'HERPES SIMPLEX, TIPO 1 e 2, ANTICORPOS IgG, soro' 'Herpes Simplex, Tipo 1 e 2, IgG'
- 'HERPES SIMPLEX, TIPO 1 e 2, ANTICORPO IgM, soro' 'Herpes Simplex, Tipo 1 e 2, IgM'
- 'CAXUMBA, ANTICORPOS IgG, soro' 'Caxumba, IgG'
- 'CAXUMBA, ANTICORPOS IgM, soro' 'Caxumba, IgM
- 'HELICOBACTER PYLORI, ANTICORPOS IgG, soro' 'Helicobacter Pylori, IgG'
- 'ERITROVIRUS B19, ANTICORPOS IgM, soro' 'Eritrovírus B19, IgM'
- 'ANTICORPOS IgM CONTRA O VIRUS DA HEPATITE A (ANTI-VHA IgM)' 'Anti-VHA, IgM'
- 'ANTICORPO IgM CONTRA ANTIGENO DO CORE DO VIRUS DA HEPATITE B, soro' 'Anti-HBc, IgM'
- 'TOXOPLASMA, ANTICORPOS IgG, soro' 'Toxoplasma, IgG'
- 'TOXOPLASMA, ANTICORPOS IgM, soro' 'Toxoplasma, IgM'

- Tem Resultados redundantes em alguns exames, por exemplo, eles estão expressados de duas maneiras, primeiro dado original, e a outra forma é de maneira normalizada (%), exemplo:

15	542	HEMOGRAMA, sangue total	Neutrófilos (%)	35484
16	541	HEMOGRAMA, sangue total	Neutrófilos	35484
17	537	HEMOGRAMA, sangue total	Monócitos	35484
18	538	HEMOGRAMA, sangue total	Monócitos (%)	35484
19	532	HEMOGRAMA, sangue total	Linfócitos (%)	35484
20	528	HEMOGRAMA, sangue total	Hemoglobina	35484
21	531	HEMOGRAMA, sangue total	Linfócitos	35484
22	519	HEMOGRAMA, sangue total	Basófilos (%)	35483
23	540	HEMOGRAMA, sangue total	Morfologia, Série Vermelha	35483
24	524	HEMOGRAMA, sangue total	Eosinófilos (%)	35483
25	518	HEMOGRAMA, sangue total	Basófilos	35483
26	523	HEMOGRAMA, sangue total	Eosinófilos	35483

- Entonces aqui deletamos o original, e mantemos o normalizado.

Segue, análise dos rótulos dos estágios da COVID19, isto é, DE_ANALITO que confirme se tem COVID19 ou que não tem, DE_ANALITO que confirme se tem IgG ou que não tem, e DE_ANALITO que confirma se tem IgM ou que não tem.

PCR

O único 'DE_ANALITO' relacionado com a confirmação de COVID19 é:

- **'Covid 19, Detecção por PCR'** tem 50223 ocorrencias.

Os 'DE_RESULTADO' encontrados no dataset são:

```
NÃO DETECTADO               19285
NÃO DETECTADO (NEGATIVO)    18287
DETECTADO                    8159
DETECTADO (POSITIVO)         4362
Inconclusivo                   48
INCONCLUSIVO                   41
Inconclusivo                   40
INCONCLUSIVO                    1
Name: DE_RESULTADO, dtype: int64
```

Deletamos os INCONCLUSIVOS, pois não se tem certeza se tem COVID19 ou não tem, além disso temos poucos registros deste tipo.

Agora temos 4 tipos de respostas redundantes, vamos converter apenas para resposta binária, onde:

- positivo/detectado: 1
- não detectado/negativo: 0

Finalmente, temos:

```
IALITO'] == 'Covid 19, Detecção por PCR']['DE_RESULTADO'].value_counts()
```

```
0    37572
1    12521
Name: DE_RESULTADO, dtype: int64
```

IgG

Temos os seguintes DE_EXAME, DE_ANALITO relacionadas con anticorpos IgG Covid19:

- 'COVID19, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Quimiolumin.-Índice'
- 'COVID19, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Quimioluminescência'

Ambos explicam a mesma coisa de duas formas diferentes, pois o índice é a quantidade de **índice** IgG, porém o outro teste tem o resultado, que nada mais que é quando aplica um umbral de 0.8 no **índice,** se ele for inferior do 0,8, o resultado deste exame é NÂO REAGENTE, caso contrário é REAGENTE. Deletamos os resultados que diz 'Indeterminado' e fica da seguinte forma

```
.9, Anticorpos IgG, Quimioluminescência']['DE_RESULTADO'].value_counts()
```

```
NÃO REAGENTE    73795
REAGENTE         7894
Name: DE_RESULTADO, dtype: int64
```

Também mapeamos esses valores para valores numéricos:

```
[resultados_exames['DE_ANALITO'] == 'Covid 19, Anticorpos IgG, Quimiolu
```

```
0    73795
1     7894
Name: DE_RESULTADO, dtype: int64
```

Também tem-se esse DE_EXAME, DE_ANALITO que obtém o IgG

- 'COVID19, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Elisa - Índice'
- 'COVID19, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Elisa'

Processando e mapeando eles para dados numéricos temos:

```
']  == 'Covid 19, Anticorpos IgG, Elisa']['DE_RESULTADO'].value_counts()
```

```
0    3301
1     543
Name: DE_RESULTADO, dtype: int64
```

Além disso, também temos este exame q calcula anticorpos IgG:

- 'SARS-CoV-2, ANTICORPOS IgM E IgG, TESTE RÁPIDO' 'Covid 19, Anticorpos IgG, teste rápido'

```
index_covid_igg = resultados_exames.loc[resultados_exames['DE_ANALITO']
index_covid_igg.replace({'REAGENTE': 1, \
                         'NÃO REAGENTE': 0 }, inplace=True)
resultados_exames.loc[index_covid_igg.index, 'DE_RESULTADO'] = index_co
resultados_exames[resultados_exames['DE_ANALITO'] == 'Covid 19, Anticor
```

```
0    2853
1     183
Name: DE_RESULTADO, dtype: int64
```

E por último:

- 'SARS-COV-2, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Elisa - Índice'
- 'SARS-COV-2, ANTICORPOS IgG, soro' 'Covid 19, Anticorpos IgG, Elisa'

```
ains('Covid 19, Anticorpos IgG, Elisa')]['DE_RESULTADO'].value_counts()
```

```
0    3301
1     543
Name: DE_RESULTADO, dtype: int64
```

Por tanto, os DE_ANALITO associados a anticorpos IgG que mais pra frente vão ser nossos rótulos para nossa rede IgG são:

- 'Covid 19, Anticorpos IgG, Quimioluminescência'
- 'Covid 19, Anticorpos IgG, Elisa'
- 'Covid 19, Anticorpos IgG, teste rápido'

IgM

DE_EXAME e DE_ANALITO associados:

- 'SARS-CoV-2, ANTICORPOS IgM E IgG, TESTE RÁPIDO' 'Covid 19, Anticorpos IgM, teste rápido'

```
index_covid_igm = resultados_exames.loc[resultados_exames['DE_ANALITO']
index_covid_igm.replace({'REAGENTE': 1, \
                    'NÃO REAGENTE': 0 }, inplace=True)
resultados_exames.loc[index_covid_igm.index, 'DE_RESULTADO'] = index_co
resultados_exames[resultados_exames['DE_ANALITO'] == 'Covid 19, Anticor
```

```
0    2859
1     177
Name: DE_RESULTADO, dtype: int64
```

Também os seguintes:

- COVID19, ANTICORPOS IgM, soro Covid 19, Anticorpos IgM, Quimiolumin.-Índice
- COVID19, ANTICORPOS IgM, soro Covid 19, Anticorpos IgM, Quimioluminescência

```
index_covid_igm = resultados_exames.loc[resultados_exames['DE_ANALITO']
index_covid_igm.replace({'REAGENTE': 1, \
                    'NÃO REAGENTE': 0 }, inplace=True)
resultados_exames.loc[index_covid_igm.index, 'DE_RESULTADO'] = index_co
resultados_exames[resultados_exames['DE_ANALITO'] == 'Covid 19, Anticor
```

```
0    73748
1     5557
Name: DE_RESULTADO, dtype: int64
```

Por tanto, os DE_ANALITO associados a anticorpos IgM que mais pra frente vão ser nossos rótulos para nossa rede IgM são:

- 'Covid 19, Anticorpos IgM, Quimioluminescência'
- 'Covid 19, Anticorpos IgM, teste rápido'

Uma vez tendo os rótulos prontos para os 3 estágios da COVID19, seguimos analisando os registros dos exames.

Análisis gerais

Deletamos as seguintes colunas, pois não agregam valor :

- ['DE_ORIGEM', 'CD_UNIDADE', 'DE_VALOR_REFERENCIA']

Se a densidade de DE_ANALITO for menor do que 400 deletamos o registro que contenha esses DE_ANALITO, pois 400 é muito pouco comparado ao número de pacientes, por tanto tem muito paciente que não se fizeram esses exames.

```
exames_agrupados[exames_agrupados['VEZES'] < 400]
```

	DE_EXAME	DE_ANALITO	VEZES
1256	VITAMINA E (tocoferol), soro	Vitamina E (tocoferol), soro	399
910	PROTEINA C-REATIVA, plasma	Proteína C-Reativa	387
244	CKMB massa, plasma	CK-MB, Eletroquimioluminescência	363
289	CREATININA, plasma	Creatinina	363
1139	TRANSFERRINA, soro	Transferrina	360
...	...	...	...
945	QUANTIFICACAO DE INIBIDOR DE FATOR IX, plasma	Dosagem de fator IX	1
1141	TRIAGEM DE ANFETAMINAS E METANFETAMINAS, urina	Anfetaminas e Metanfetaminas, urina	1
759	MAG, ANTICORPOS, soro	MAG, anticorpos	1
852	PESQUISA DE MICROSPORIDIA NAS FEZES	Microsporidia, pesquisa, fezes	1
731	JO-1, AUTO-ANTICORPOS, soro	Jo-1, Auto-Anticorpos - imunodifusão dupla	1

1014 rows × 3 columns

Outra coisa que se avaliou foi a quantidade de EXAMES FEITO POR PACIENTE:

```
resultados_exames['ID_PACIENTE'].value_counts()
```

```
CE1F4D06FE83023E08DF680EC7324BF5    649
A9DEF7F65875D4B92E59D50F9B21DCF2    644
FD10C24CA5986B00FFDDED743447EC81    643
k to scroll output; double click to hide 4CE    587
FC5675885071BBF89EE97CFFBB80B5D8    558
                                   ...
D5777D57B14E3CB2E2ED600C2EC86742      1
B24C127D85F494A6145A60EE30246FE7      1
C01D8355ABF56655EB5989E8789E663A      1
5BFE3D28F619425E2E25D160F267E198      1
0F4844A9976A99DC30DE8E011C0BA4DF      1
Name: ID_PACIENTE, Length: 128706, dtype: int64
```

Só vamos analisar pacientes que tenham feito como mínimo 18 exames DE_ANALITO:

```
resultados_exames['ID_PACIENTE'].value_counts()
```

```
CE1F4D06FE83023E08DF680EC7324BF5    649
A9DEF7F65875D4B92E59D50F9B21DCF2    644
FD10C24CA5986B00FFDDED743447EC81    643
4AD9371C8FAF9B4B3C409589E09104CE    587
FC5675885071BBF89EE97CFFBB80B5D8    558
                                   ...
A420012B68C72D4CD71088AD560A302F     18
4EFF38F5EC982B7A9F1AEFB154154A3C     18
C2060E3D27DC2CAEF2A28E14E108061D     18
3DBB1C0ADD2799027ED193569CE8B4C1     18
AFBA550AA6D3DE60D29DB92B431AC854     18
Name: ID_PACIENTE, Length: 26032, dtype: int64
```

Propagação de resultados dos testes por ID_PACIENTE e DATA_COLETA

Agrupamos por paciente, e ordenamos segundo a DATA_COLETA, e utilizamos o DE_RESULTADO da última DATA_COLETA que o paciente fez o exame de interesse. O resultado da propagração esta no arquivo: **resultados_pcr_prograsao.csv**
E os rotulos que vamos utilizar por cada ID_PACIENTE estão salvos no arquivo: **paciente_covid_resultado_final.csv**

Arquitetura da rede neural

Minha proposta é treinar uma rede por cada ESTÁGIO DA COVID19

CNN PCR

As features para esta Rede são os top 30 DE_ANALITOS que são os exames que tem resultado de COVID19 positivo ou não para serem utilizados como label no momento de treinar e testar a rede.

```
resultados_exames_covid19.DE_ANALITO.value_counts()[:30]
```

```
Hematócrito                                11927
Concentração de Hemoglobina Corpuscular    11927
Leucócitos                                 11927
Hemoglobina Corpuscular Média              11927
Linfócitos (%)                             11927
Plaquetas                                  11927
Monócitos (%)                              11927
Eritrócitos                                11927
VCM                                        11927
Neutrófilos (%)                            11927
Hemoglobina                                11926
Morfologia, Série Branca                   11926
Basófilos (%)                              11926
Eosinófilos (%)                            11926
Morfologia, Série Vermelha                 11926
RDW                                        11925
Volume plaquetário médio                   11662
Covid 19, Detecção por PCR                  9117
Creatinina                                  8986
Glicose                                     8462
ALT (TGP)                                   8359
AST (TGO)                                   8283
Uréia                                       7877
Hormônio Tiroestimulante                    7521
Triglicérides                               7173
Colesterol total                            7141
HDL-Colesterol                              7097
LDL-Colesterol                              6955
Colesterol não-HDL, soro                    6948
VLDL-Colesterol                             6699
Name: DE_ANALITO, dtype: int64
```

Note-se que temos 9117 rótulos, isto quer dizer que teremos 29 colunas que são os exames restantes e 1 coluna como o label associado ao registro de resultados dos exames. No momento temos o arquivo sem ainda converter no formato de colunas de features, que é **resultados_exames_covid19.csv**

CNN IgG

CNN IgM

Descrição dos experimentos

Resultados

Discussão

Bibliografia