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

Frederico Comério

O IMPACTO DESIGUAL DA PANDEMIA DE COVID-19 NA SOCIEDADE BRASILEIRA

Belo Horizonte
2021

Frederico Comério

O IMPACTO DESIGUAL DA PANDEMIA DE COVID-19 NA SOCIEDADE BRASILEIRA

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2021

SUMÁRIO

1. Introdução

1.1. Contextualização

A pandemia de COVID-19 vem mobilizando todos os setores da sociedade em busca de entendimento, conscientização e alternativas para identificação de incidência, rastreamento da expansão do contágio, entre tantas outras tantas análises. As ciências, em todas as suas vertentes, encontram-se empenhadas nas pesquisas e observações acerca deste evento, com destaque para a de Ciência de Dados.

Segundo ORACLE (2020), a ciência de dados combina vários campos, incluindo estatísticas, ciência da comutação, métodos científicos e análise de dados para extrair valor dos dados. Aqueles que praticam a ciência de dados são chamados de cientistas de dados e combinam uma variedade de habilidades para analisar dados coletados de diversas fontes como web, smartphones, clientes, sensores de IOT, entre outros.

De acordo com AQUARELA (2020), a Ciência de dados é um campo interdisciplinar de investigação de dados que resolve problemas reais de negócios, com o uso de método científico e técnicas avançadas de análise de dados, aprendizado de máquina e inteligência artificial.

Este trabalho visa utilizar de técnicas e ferramentas de ciência de dados para analisar e ilustrar como os dados de óbitos do COVID-19 explicam impactos e desigualdades sociais dentro do município de São Paulo e que servem de amostragem para prever como esta análise pode refletir uma situação análoga no país como um todo.

1.2. O problema proposto

Os dados de óbitos decorrentes da pandemia de COVID-19 não traduzem de maneira clara os diferentes impactos em diferentes camadas da sociedade. A análise exploratória conduzida neste trabalho utiliza de síntese estatística, engenharia de requisitos (*feature engineering*) e técnicas de visualização de dados para melhor compreensão das informações e identificação de insights e tendências dentro do contexto da pandemia, bem como formular e validar hipóteses dentro do contexto dos problemas sociais.

Para maior elucidação, segue uma descrição dos principais aspectos do problema a ser resolvido utilizando a técnica dos 5-Ws:

a) Por que esse problema é importante (*Why?*)

A pandemia de COVID-19 é provavelmente o assunto mais relevante dos últimos meses, e a desigualdade social é um dos mais antigos problemas do país. Entender como esses eventos se correlacionam e seus reais impactos são de suma importância para o tratamento de ambos os problemas.

b) De quem são os dados analisados? (*Who?*)

Os dados analisados são provenientes da Secretaria Municipal de Saúde de São Paulo a partir do sistema TABNET (instrumento que possibilita o acesso às bases de dados de população e dos sistemas de informações do SUS).

c) Quais os objetivos com essa análise? (*What?*)

Entender a partir de diferentes fontes de dados (datasets) como a pandemia de COVID-19 impacta de maneira desigual grupos mais ou menos vulneráveis da sociedade.

d) Local de estudo (*Where?*):

Indivíduos residentes no município de São Paulo.

e) Qual o período está sendo analisado? (*When?*)

Ano de 2020.

2. Coleta de Dados

Conforme descrito no capítulo anterior, os dados analisados são provenientes da Secretaria Municipal de Saúde de São Paulo a partir do sistema TABNET (instrumento que possibilita o acesso às bases de dados de população e dos sistemas de informações do SUS).

De acordo com PREFEITURA DE SÃO PAULO, 2021, A Secretaria Municipal da Saúde de São Paulo disponibiliza o TABNET, instrumento que possibilita o acesso às bases de dados de população e dos sistemas de informações do SUS como: mortalidade, nascimentos, procedimentos ambulatoriais, internações hospitalares, estabelecimentos de saúde, saúde da família, doenças, imunização, acidentes de trabalho e acidentes.

Mais recentemente, foram acrescentadas bases de dados como de algumas doenças e agravos de notificação compulsória, profissionais ativos na SMS e do ISA-Capital, inquérito de saúde de base populacional, entre outros.

O TABNET é um aplicativo web desenvolvido pelo DATASUS que permite extração de dados a partir de tabulações cruzando-se diversas variáveis segundo o interesse do usuário. As bases de dados são atualizadas periodicamente.

As principais fontes de dados utilizados neste trabalho estão descritas na tabela abaixo e compartilhados em:

https://github.com/fredcobain/tcc_pos_data_pucmg/tree/master/dados_2021/originais_datasus :

Nome do Arquivo:	Descrição:	Fonte:	Download em:
cor-se-suspeitos.csv	Óbitos de residentes no município SP por Cor e semana epidemiológica no período de 2020.	http://tabnet.saude.prefeitura.sp.gov.br/cgi/deftohtm3.exe?secretarias/saude/TABNET/SIM_PROV/obitop.def	04/07/2021
cor-idade-pop2010.csv	População residente segundo sexo, faixa etária, raça / cor no município de São Paulo no último censo realizado no ano de 2010.	http://tabnet.saude.prefeitura.sp.gov.br/cgi/deftohtm3.exe?secretarias/saude/TABNET/POPRC/poprc.def	04/07/2021
cor-idade-suspeitos-2020.csv	Óbitos confirmados por COVID-19 no município de São Paulo no período de 2020 agrupados por cor e faixa etária.	http://tabnet.saude.prefeitura.sp.gov.br/cgi/deftohtm3.exe?secretarias/saude/TABNET/SIM_PROV/obitop.def	04/07/2021

Tabela 1: Principais Fontes de Dados utilizadas

As tabelas abaixo ilustram detalhes acerca da estrutura de campos dos respectivos datasets:

Nome da coluna/campo	Descrição	Tipo
Cor	Cor autodeclarada dos indivíduos.	Texto.
SE_20_11 ... SE_20_52	Semana do ano de 2021	Numérico.

Tabela 2: Estrutura do dataset cor-se-suspeitos.csv

Cor	SE_2011_1ºÓbito Covid19	SE_2012	SE_2013	SE_2014	SE_2015	SE_2016	SE_2017	SE_2018	SE_2019	...	SE_2045	SE_2046	SE_2047	SE_2048	SE_2049	SE_2050
Branca	3	48	178	362	437	442	501	634	615	...	105	137	159	192	228	262
Preta	-	2	15	54	64	55	78	102	86	...	14	16	16	30	27	40
Amarela	-	2	10	16	17	19	28	24	30	...	2	6	10	9	8	11
Parda	-	5	37	126	171	187	207	283	301	...	36	42	48	51	76	66
Indígena	-	-	1	-	2	-	1	-	2	...	1	-	-	-	-	1
Não informado	-	-	12	17	25	28	22	41	41	...	3	3	1	7	8	15
Total	3	57	253	575	716	731	837	1084	1075	...	161	204	234	289	347	395

Figura 1: Visualização do dataset cor-se-suspeitos.csv

Nome da coluna/campo	Descrição	Tipo
Cor	Cor autodeclarada dos indivíduos.	Texto.
0 a 4 ... 75 e mais	Faixa etária dos indivíduos	Numérico.

Tabela 3: Estrutura do dataset cor-idade-pop2010.csv

Raça / Cor	0 a 4	5 a 9	10 a 14	15 a 19	20 a 24	25 a 29	30 a 34	35 a 39	40 a 44	45 a 49	50 a 54	55 a 59	60 a 64	65 a 69
Branca	444168	433860	466843	462715	570950	636194	599762	524122	485640	465774	425601	358120	283006	208206
Preta	29954	39476	52706	56272	68610	75869	73139	64755	59093	51910	47914	37809	27559	18537
Amarela	8329	9020	10726	12276	17282	21097	20445	18141	17467	17002	17422	17874	17175	13987
Parda	227650	275068	336289	309997	333423	339887	315436	280563	249819	207196	175970	133704	94896	61310
Indígena	813	840	843	982	1352	1484	1256	1072	939	823	734	590	413	293
Ignorado	13	15	23	15	42	52	38	31	21	15	17	16	6	5
Total	710927	758279	867430	842257	991659	1074583	1010076	888684	812979	742720	667658	548113	423055	302338

Figura 2: Visualização do dataset cor-idade-pop2010.csv

Nome da coluna/campo	Descrição	Tipo
Cor	Cor autodeclarada dos indivíduos.	Texto.
0 a 4a ... 75 e mais	Faixa etária dos indivíduos	Numérico.

Tabela 4: Estrutura do dataset cor-idade-suspeitos.csv

Cor	0-4a	5-9a	10-14a	15-19a	20-24a	25-29a	30-34a	35-39a	40-44a	45-49a	50-54a	55-59a	60-64a	65-69a	70-74a	75 e mais	Ign	Total
Branca	18	8	8	17	41	52	125	177	255	318	528	825	1200	1504	1736	7482	3	14297
Preta	-	1	-	4	8	11	24	49	68	92	119	173	235	280	277	765	2	2108
Amarela	-	-	-	1	-	-	-	2	2	8	10	25	28	42	65	361	-	544
Parda	17	5	5	13	27	43	87	139	197	254	390	437	616	695	693	1747	2	5367
Indígena	1	-	1	-	-	-	1	-	-	1	-	1	-	2	2	7	-	16
Não informado	2	-	-	1	5	-	3	12	18	27	40	67	77	88	84	331	1	756
Total	38	14	14	36	81	106	240	379	540	700	1087	1528	2156	2611	2857	10693	8	23088

Figura 3: Visualização do dataset cor-idade-suspeitos.csv

Segundo a SECRETARIA MUNICIPAL DA SAÚDE DE SÃO PAULO, o método utilizado para classificação da coluna COR segue o sistema classificatório

do IBGE (OSORIO, 2003), ou seja, método da AUTOCLASSIFICAÇÃO ou AUTODECLARAÇÃO. Neste método, o indivíduo é quem indica a sua “cor ou raça/etnia” entre as cinco categorias possíveis: branca, preta, parda, amarela, indígena.

Haverá situações em que será necessário utilizar a heteroclassificação, isto é, outra pessoa, preferencialmente um membro da família, define a cor ou raça/etnia do indivíduo, mas esta conduta deverá ser utilizada somente em situações específicas, tais como: declaração de nascidos vivos, declaração de óbito, registro de pacientes em coma ou quadros semelhantes (DIAS, 2009).

Este método também possibilita o cruzamento dos dados obtidos em todo o país. Assim, é possível fazer comparações abrangentes e ter estatísticas em nível nacional.

Além disso, antes de definir estas categorias, o IBGE pesquisou as cores mais declaradas pela população e concluiu que deveria utilizar este conjunto, pois a maioria destas já são utilizadas em pesquisas nacionais desde a segunda metade do século XIX (o que permite análises históricas destes dados e categorias).

Por fim, foram selecionados os dados específicos do município de São Paulo para esta análise não só por conta da riqueza e qualidade dos registros, mas também por ser um município referência no país em termos de amostragem, pois, além de ser a maior cidade do Brasil, São Paulo é a cidade mais populosa da América do Sul (e de todo o hemisfério sul com) mais de 12 milhões de habitantes (o que representa aproximadamente 6% da população nacional).

3. Processamento e Tratamento de Dados

3.1 Ferramentas e Tecnologias utilizadas

Na etapa de processamento e tratamento de dados (e também no processo de análise exploratória) foram realizadas com apoio das seguintes tecnologias, frameworks e ferramentas:

- Linguagem de Programação: Python 3.8

- Distribuição: Anaconda Framework 4.10.1

- IDE: Visual Studio Code 1.52

- Extensão: Jupyter Notebook v2021.5.745244803

A linguagem Python foi escolhida por ser uma linguagem extremamente flexível, de código aberto e amplamente utilizada para manipulação de dados. Suas incontáveis bibliotecas são de rápida curva de aprendizado e bem documentadas.

A distribuição Anaconda foi escolhida por possuir as principais ferramentas e bibliotecas para criação de scripts de análise de dados em seu pacote, evitando assim a necessidade de instalação e configuração manual de bibliotecas tradicionais de ciência de dados para a linguagem Python, otimizando o desenvolvimento dos scripts.

O editor VSCode, foi utilizado como ferramenta de edição de código e depuração por ser gratuito altamente customizável e com um rico conjunto de extensões que podem ser usadas para melhorar as funcionalidades padrão.

A extensão Jupyter Notebook para VSCode foi escolhida por ser uma ferramenta de Literate Computing extremamente eficiente, a qual permite unir código e texto em diferentes células de processamento. Desta forma, cada trecho ou funcionalidade pode ser documentada e processada separadamente.

3.2 Limpeza e Tratamento no Dataset 1 - cor-se-suspeitos.csv

O dataset “cor-se-suspeitos.csv”, identificado neste trabalho como “Dataset 1”, foi adquirido através do sistema TABNET (vide capítulo 2) e possui originalmente uma série de linhas informativas que já podem ser descartadas no momento do seu carregamento:

1	Mortalidade geral exceto causas externas									
2	Óbitos Residentes MSP por Cor e Semana epidemiológica									
3	Causas específicas: Óbitos suspeitos de Covid 19, Óbitos confirmados de Covid 19									
4	Período:2020									
5	Cor	SE_20 11_	SE_20 12	SE_20 13	SE_20 14	SE_20 15	SE_20 16	SE_20 17	SE_20 18	SE_20 19	SE_20 20
6	Branca	3	48	178	362	437	442	501	634	615	685
7	Preta	-	2	15	54	64	55	78	102	86	120
8	Amarela	-	2	10	16	17	19	28	24	30	18
9	Parda	-	5	37	126	171	187	207	283	301	298
10	Indígena	-	-	1	-	2	-	1	-	2	1
11	Não inforr	-	-	12	17	25	28	22	41	41	40
12	Total	3	57	253	575	716	731	837	1084	1075	1162
13	Fonte: Sistema de Informações sobre Mortalidade – SIM/PRO-AIM/CEInfo – SMS/SP. Data de atualização: 02/07/2021.										
14	Nota:										
15	1- Para tabulações de proporções, o campo referente à proporção deve constar em linhas ou colunas.										
16	2- Em acordo com orientação do Ministério da Saúde de 20/03/2020, o código U04.9 foi utilizado como marcador dos ca										

Figura 4: Seção do Dataset 1 em formato raw

Desta forma, a importação do dataset numa estrutura de dados conhecida como “PANDAS” ocorreu ignorando as linhas 0,1,2,3,12,13,14,15, as quais não havia dados de fato através do comando abaixo:

```
cor_se_temp = pd.read_csv('cor-se-suspeitos.csv', sep=';', encoding = "ISO-
8859-1",skiprows=[0,1,2,3,12,13,14,15])
```

Com isso, apenas os dados úteis foram carregados de forma que a estrutura de dados ficou apresentada da seguinte maneira:

Cor	SE_20 11_1ºÓbito Covid19	SE_20 12	SE_20 13	SE_20 14	SE_20 15	SE_20 16	SE_20 17	SE_20 18	SE_20 19	...	SE_20 45	SE_20 46	SE_20 47	SE_20 48	SE_20 49	SE_20 50	SE_20 51	SE_20 52	SE_20 53
Branca	3	48	178	362	437	442	501	634	615	...	105	137	159	192	228	262	301	284	191
Preta	-	2	15	54	64	55	78	102	86	...	14	16	16	30	27	40	39	44	27
Amarela	-	2	10	16	17	19	28	24	30	...	2	6	10	9	8	11	11	15	15
Parda	-	5	37	126	171	187	207	283	301	...	36	42	48	51	76	66	84	80	64
Indígena	-	-	1	-	2	-	1	-	2	...	1	-	-	-	-	1	-	-	-
Não informado	-	-	12	17	25	28	22	41	41	...	3	3	1	7	8	15	16	11	9
Total	3	57	253	575	716	731	837	1084	1075	...	161	204	234	289	347	395	451	434	306

Figura 5: Dataset 1 importado para a estrutura de PANDAS

Na figura acima, é possível identificar uma série de campos numéricos com valores do tipo “-”, onde na verdade é necessário atualizar para o valor numérico “0” a fim de que todo o campo de valores seja calculável, vide o comando abaixo:

```
cor_se = cor_se_temp.apply(pd.to_numeric, errors='coerce').fillna(0)
```

O que resulta para a seguinte estrutura de dataset:

Cor	SE_20 11_1ºÓbito Covid19	SE_20 12	SE_20 13	SE_20 14	SE_20 15	SE_20 16	SE_20 17	SE_20 18	SE_20 19	...	SE_20 45	SE_20 46	SE_20 47	SE_20 48	SE_20 49	SE_20 50	SE_20 51	SE_20 52
0.0	3.0	48.0	178	362.0	437	442.0	501	634.0	615	...	105	137.0	159.0	192.0	228.0	262	301.0	284.0
0.0	0.0	2.0	15	54.0	64	55.0	78	102.0	86	...	14	16.0	16.0	30.0	27.0	40	39.0	44.0
0.0	0.0	2.0	10	16.0	17	19.0	28	24.0	30	...	2	6.0	10.0	9.0	8.0	11	11.0	15.0
0.0	0.0	5.0	37	126.0	171	187.0	207	283.0	301	...	36	42.0	48.0	51.0	76.0	66	84.0	80.0
0.0	0.0	0.0	1	0.0	2	0.0	1	0.0	2	...	1	0.0	0.0	0.0	0.0	1	0.0	0.0
0.0	0.0	0.0	12	17.0	25	28.0	22	41.0	41	...	3	3.0	1.0	7.0	8.0	15	16.0	11.0
0.0	3.0	57.0	253	575.0	716	731.0	837	1084.0	1075	...	161	204.0	234.0	289.0	347.0	395	451.0	434.0

Figura 6: Dataset 1 sem valores textuais

Obviamente, a coluna cor precisa ser restaurada com os valores textuais iniciais com o comando abaixo:

```
cor_se['Cor'] = cor_se_temp['Cor']
```

Da mesma forma, a coluna da primeira semana epidemiológica está nominada num padrão diferente das demais, o que pode ser corrigido com o comando abaixo:

```
cor_se.rename({'SE_20 11_1ºÓbito Covid19': 'SE_20 11'}, axis=1, inplace=True)
display(cor_se)
```

A intenção com esse dataset é observar a soma cumulativa através das semanas, e não o número de óbitos em cada semana. Para isto, o método “cumsum” pode ser aplicado para realizar este cálculo:

```
cor_se_norm = cor_se_norm.cumsum(axis=1)
```

O que resulta na estrutura abaixo:

SE_2011	SE_2012	SE_2013	SE_2014	SE_2015	SE_2016	SE_2017	SE_2018	SE_2019	SE_2020	...	SE_2045	SE_2046	SE_2047	SE_2048	SE_2049	SE_2050
3.0	51.0	229.0	591.0	1028.0	1470.0	1971.0	2605.0	3220.0	3905.0	...	12543.0	12680.0	12839.0	13031.0	13259.0	13521.0
0.0	2.0	17.0	71.0	135.0	190.0	268.0	370.0	456.0	576.0	...	1869.0	1885.0	1901.0	1931.0	1958.0	1998.0
0.0	2.0	12.0	28.0	45.0	64.0	92.0	116.0	146.0	164.0	...	459.0	465.0	475.0	484.0	492.0	503.0
0.0	5.0	42.0	168.0	339.0	526.0	733.0	1016.0	1317.0	1615.0	...	4856.0	4898.0	4946.0	4997.0	5073.0	5139.0
0.0	0.0	1.0	1.0	3.0	3.0	4.0	4.0	6.0	7.0	...	15.0	15.0	15.0	15.0	15.0	16.0
0.0	0.0	12.0	29.0	54.0	82.0	104.0	145.0	186.0	226.0	...	686.0	689.0	690.0	697.0	705.0	720.0
3.0	60.0	313.0	888.0	1604.0	2335.0	3172.0	4256.0	5331.0	6493.0	...	20428.0	20632.0	20866.0	21155.0	21502.0	21897.0

Figura 7: Dataset 1 com soma cumulativa

Desta forma, o primeiro tratamento de dados no conjunto de dados “cor-se-suspeitos.csv” foi finalizado e o resultado salvo no arquivo “cor-semana-covid-tratado.csv” através do comando:

```
cor_se_norm.to_csv('cor-semana-covid-tratado.csv')
```

Obs: Todo o código e tratamento de dados realizado nesta seção (3.2) está disponível no arquivo:

https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/01%20-%20cor-semana.ipynb

3.3 Limpeza e Tratamento no Dataset 2 - cor-idade-pop2010.csv

O dataset “cor-idade-pop2010.csv”, identificado neste trabalho como “Dataset 2”, foi adquirido através do sistema TABNET (vide capítulo 2), e possui originalmente uma série de linhas informativas que já podem ser descartadas no momento do seu carregamento:

1	População residente segundo sexo, faixa etária, raça / cor e local de residência. Município de São Paulo										
2	Populacao por Raça / Cor e Faixa Etária										
3	Período:2010										
4	Raça / Cor	0 a 4	5 a 9	10 a 14	15 a 19	20 a 24	25 a 29	30 a 34	35 a 39	40 a 44	45 a 49	50 a 54
5	Branca	444168	433860	466843	462715	570950	636194	599762	524122	485640	465774	425601
6	Preta	29954	39476	52706	56272	68610	75869	73139	64755	59093	51910	47914
7	Amarela	8329	9020	10726	12276	17282	21097	20445	18141	17467	17002	17422
8	Parda	227650	275068	336289	309997	333423	339887	315436	280563	249819	207196	175970
9	Indígena	813	840	843	982	1352	1484	1256	1072	939	823	734
10	Ignorado	13	15	23	15	42	52	38	31	21	15	17
11	Total	710927	758279	867430	842257	991659	1074583	1010076	888684	812979	742720	667658
12	Fonte: Censo demográfico (IBGE), 2010											

Figura 8: Seção do Dataset 2 em formato RAW

Desta forma, a importação do dataset numa estrutura de dados conhecida como “PANDAS” ocorreu ignorando as linhas 0,1,2,11, as quais não havia dados de fato através do comando abaixo:

```
pop_cor_idade_temp = pd.read_csv('cor-idade-
pop2010.csv', sep=';', encoding = "ISO-8859-1",
skiprows=[0,1,2,11])
```

Com isso, apenas os dados úteis foram carregados de forma que a estrutura de dados ficou apresentada da seguinte maneira (13 primeiras colunas):

Raça / Cor	0 a 4	5 a 9	10 a 14	15 a 19	20 a 24	25 a 29	30 a 34	35 a 39	40 a 44	45 a 49	50 a 54	55 a 59
Branca	444168	433860	466843	462715	570950	636194	599762	524122	485640	465774	425601	358120
Preta	29954	39476	52706	56272	68610	75869	73139	64755	59093	51910	47914	37809
Amarela	8329	9020	10726	12276	17282	21097	20445	18141	17467	17002	17422	17874
Parda	227650	275068	336289	309997	333423	339887	315436	280563	249819	207196	175970	133704
Indígena	813	840	843	982	1352	1484	1256	1072	939	823	734	590
Ignorado	13	15	23	15	42	52	38	31	21	15	17	16
Total	710927	758279	867430	842257	991659	1074583	1010076	888684	812979	742720	667658	548113

Figura 9: Dataset 2 importado para a estrutura de PANDAS

O mesmo tratamento para converter os caracteres “-” em valores numéricos (do tipo 0) realizado no dataset do capítulo anterior também foi realizado neste dataset.

As 2 últimas colunas (Ignorada e Total) foram removidas através do comando:

```
pop_norm = pop_norm[:-2] #ultima (total e ignorados)
```

A linha Ignorada foi removida através do comando:

```
pop_norm = pop_norm.drop('Ignorada', axis=1)
```

Assim, o primeiro tratamento de dados no conjunto de dados “cor-idade-pop2010.csv” foi finalizado e o resultado salvo no arquivo “cor-idade-2010-tratado.csv” através do comando:

```
pop_cor_idade.to_csv('cor-idade-2010-tratado.csv')
```

Obs: Todo o código e tratamento de dados realizado nesta seção (3.3) estão disponíveis no arquivos:

https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/02%20-%20cor-idade.ipynb

e parte em

https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/04%20-%20pop-idade.ipynb

3.4 Limpeza e Tratamento no Dataset 3 - cor-idade-suspeitos-2020.csv

O dataset “cor-idade-suspeitos-2020.csv”, identificado neste trabalho como “Dataset 3”, foi adquirido através do sistema TABNET (vide capítulo 2) e possui originalmente uma série de linhas informativas que já podem ser descartadas no momento do seu carregamento:

1	Mortalidade geral exceto causas externas									
2	Óbitos Residentes MSP por Cor e Fx etária de 5 em 5 anos									
3	Causas específicas: Óbitos suspeitos de Covid 19, Óbitos confirmados de Covid 19									
4	Período:2020									
5	Cor	0-4a	5-9a	10-14a	15-19a	20-24a	25-29a	30-34a	35-39a	40-44a	45-49a
6	Branca	18	8	8	17	41	52	125	177	255	318
7	Preta	-	1	-	4	8	11	24	49	68	92
8	Amarela	-	-	-	1	-	-	-	2	2	8
9	Parda	17	5	5	13	27	43	87	139	197	254
10	Indígena	1	-	1	-	-	-	1	-	-	1
11	Não inforr	2	-	-	1	5	-	3	12	18	27
12	Total	38	14	14	36	81	106	240	379	540	700
13	Fonte: Sistema de Informações sobre Mortalidade – SIM/PRO-AIM/CEInfo – SMS/SP. Data de atualização: 02/07/2021.										
14	Nota:										
15	1- Para tabulações de proporções, o campo referente à proporção deve constar em linhas ou colunas.										
16	2- Em acordo com orientação do Ministério da Saúde de 20/03/2020, o código U04.9 foi utilizado como marcador dos ca										
17											

Figura 10: Seção do Dataset 3 em formato RAW

Desta forma, a importação do dataset numa estrutura de dados conhecida como “PANDAS” ocorreu ignorando as linhas 0,1,2,3,12,13,14,15, as quais não havia dados de fato através do comando abaixo:

```
cor_idade_temp = pd.read_csv('cor-idade-suspeitos-
2020.csv', sep=';', encoding = "ISO-8859-1",skiprows=[0,1,2,3,12,13,14,15])
```

Com isso, apenas os dados úteis foram carregados de forma que a estrutura de dados ficou apresentada da seguinte maneira

Cor	0-4a	5-9a	10-14a	15-19a	20-24a	25-29a	30-34a	35-39a	40-44a	45-49a	50-54a	55-59a	60-64a	65-69a	70-74a	75 e mais	Ign	Total
Branca	18	8	8	17	41	52	125	177	255	318	528	825	1200	1504	1736	7482	3	14297
Preta	-	1	-	4	8	11	24	49	68	92	119	173	235	280	277	765	2	2108
Amarela	-	-	-	1	-	-	-	2	2	8	10	25	28	42	65	361	-	544
Parda	17	5	5	13	27	43	87	139	197	254	390	437	616	695	693	1747	2	5367
Indígena	1	-	1	-	-	-	1	-	-	1	-	1	-	2	2	7	-	16
Não informado	2	-	-	1	5	-	3	12	18	27	40	67	77	88	84	331	1	756
Total	38	14	14	36	81	106	240	379	540	700	1087	1528	2156	2611	2857	10693	8	23088

Figura 11: Dataset 3 importado para a estrutura de PANDAS

O mesmo tratamento para converter os caracteres “-” em valores numéricos (do tipo 0) realizado no dataset do capítulo anterior também foi realizado neste dataset.

Após isso, a linha “Não informado” foi removida do dataset através do comando abaixo:

```
cor_idade = cor_idade.drop(5)
```

Assim, o primeiro tratamento de dados no conjunto de dados “cor-idade-suspeitos-2020.csv” foi finalizado.

Obs: Todo o código e tratamento de dados realizado nesta seção (3.4) está disponível no arquivo:

https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/05%20-%20mortalidade-covid-idade.ipynb

4. Análise e Exploração dos Dados

4.1 Descobrindo a taxa de mortalidade por cor/raça/etnia

Analisando de maneira crua os dados do “Dataset 1” após o processo de limpeza e processamento de dados, especialmente após aplicação do algoritmo de soma cumulativa, observa-se os resultados preliminares abaixo:

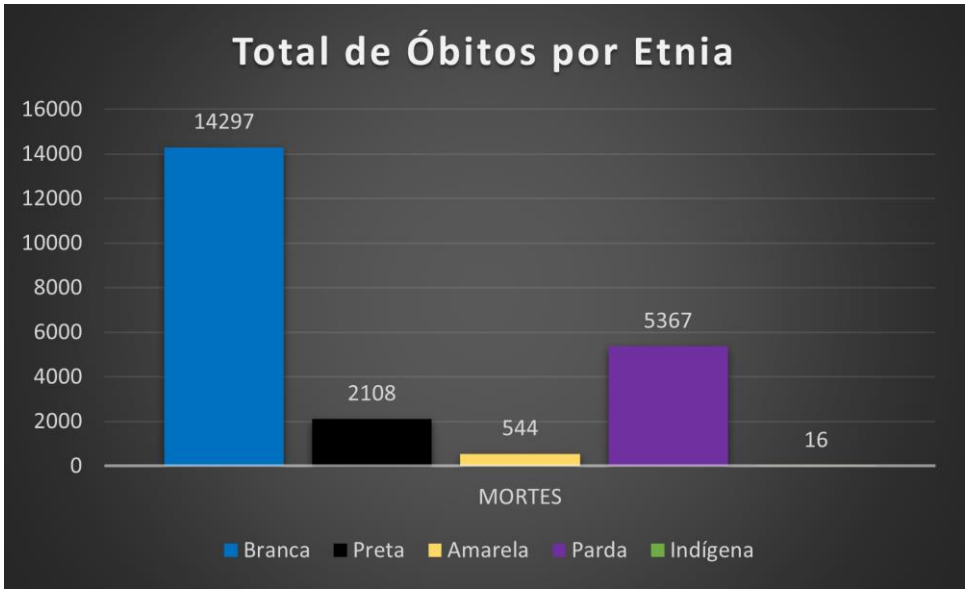

Gráfico 1: Análise preliminar do número de óbitos de COVID-19 por etnia

Obviamente isto não significa que os indivíduos autodenominados como “brancos” estão mais vulneráveis, pois o número de óbitos não leva em conta a proporção do número de habitantes que compões cada grupo étnico.

Para começar a identificar as assimetrias as quais a pandemia de COVID-19 vem causando em contextos de desigualdades entre os vários grupos étnicos, será realizado um “*merge*” entre esses dados para obtermos o índice de mortalidade a cada 100 mil habitantes por grupo étnico. Essa técnica em ciência de dados pode ser nomeada como “feature engineering”, que resumidamente seria o processo de usar o conhecimento de domínio para extrair novas variáveis a partir de uma ou mais fontes de dados.

Como o “Dataset 2” possui a contagem de habitantes por raça/cor/etnia e o “Dataset 1” possui a quantidade de óbitos por semana por raça/cor/etnia, o primeiro passo é de mover a coluna “Total” do “Dataset 2” para o “Dataset 1” através do comando:

```
cor_se_soma['Habitantes'] = pop_norm.loc[:,'Total']
```

Desta forma, a estrutura do “Dataset 1” apresenta-se como na figura abaixo:

...	SE_2046	SE_2047	SE_2048	SE_2049	SE_2050	SE_2051	SE_2052	SE_2053	Cor	Habitantes
...	12680.0	12839.0	13031.0	13259.0	13521.0	13822.0	14106.0	14297.0	Branca	6824668
...	1885.0	1901.0	1931.0	1958.0	1998.0	2037.0	2081.0	2108.0	Preta	736083
...	465.0	475.0	484.0	492.0	503.0	514.0	529.0	544.0	Amarela	246244
...	4898.0	4946.0	4997.0	5073.0	5139.0	5223.0	5303.0	5367.0	Parda	3433218
...	15.0	15.0	15.0	15.0	16.0	16.0	16.0	16.0	Indígena	12977

Figura 13: “Dataset 1” com dados de quantidade de habitantes

O próximo passo é de realizar uma transformação através do método “melt” da biblioteca “Pandas”. Este método basicamente permite transformar colunas específicas em linhas enquanto deixa outras colunas intactas. No caso, a transformação ocorrerá com as colunas “Cor” e “Habitantes”, que foi recentemente atachada no “Dataset 1”. Esta operação será realizada através do comando abaixo:

```
cor_se_melt = cor_se_soma.melt(id_vars=['Cor', 'Habitantes'], value_vars = cor
_se_norm.columns[1:-1],var_name='Semana', value_name='Óbitos')
```

Com isto, adquire-se a seguinte estrutura (considerando apenas as 5 primeiras linhas):

Cor	Habitantes	Semana	Óbitos
Branca	6824668	SE_20 53	14297.0
Preta	736083	SE_20 53	2108.0
Amarela	246244	SE_20 53	544.0
Parda	3433218	SE_20 53	5367.0
Indígena	12977	SE_20 53	16.0

Figura 14: “Dataset 1” transformado

Segundo INSTITUTO POLIS (2019), para construir leituras mais representativas da realidade desigual que a epidemia revela – e agrava – é imprescindível a padronização das taxas de mortalidade.

Nesta estrutura, fica fácil calcular o número de óbitos para cada 100 mil habitantes em cada linha, apenas adicionando uma nova coluna que será um cálculo em cima das colunas “Óbitos” e “Habitantes” através do comando abaixo:

```
cor_se_melt['Óbitos por 100 mil'] = (100000 * cor_se_melt['Óbitos']) / cor_se_
melt['Habitantes']
```

Por fim, resulta-se a estrutura definitiva para este objetivo parcial:

Cor	Habitantes	Semana	Óbitos	Óbitos por 100 mil
Branca	6824668	SE_20 53	14297.0	209.490044
Preta	736083	SE_20 53	2108.0	286.380748
Amarela	246244	SE_20 53	544.0	220.919088
Parda	3433218	SE_20 53	5367.0	156.325640
Indígena	12977	SE_20 53	16.0	123.295060

Figura 15: “Dataset 1” transformado com a taxa de óbitos/100 mil habitantes

A estrutura do dataset resultante acima permite uma análise gráfica sobre a quantidade de óbitos a cada 100 mil habitantes durante as semanas epidemiológicas do ano de 2020.

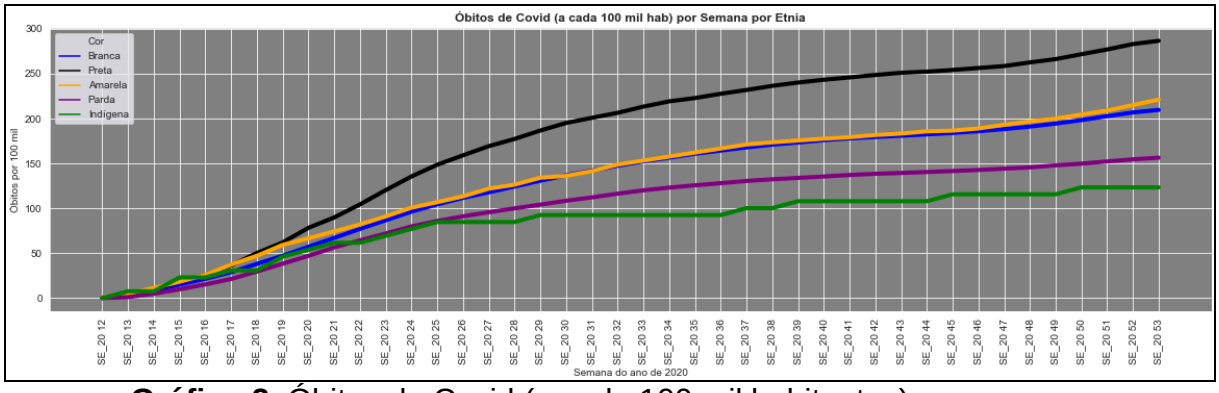

Gráfico 2: Óbitos de Covid (a cada 100 mil habitantes) por semana

Na projeção acima, observamos uma curva acentuada de indivíduos classificados como negros em detrimento dos demais grupos, o que já traz à tona uma assimetria característica da população brasileira a qual segundo IBGE(2019), a população negra 75% entre os 10% mais pobres, o que os coloca em uma escala de vulnerabilidade maior que pode ser constatada a partir do gráfico acima (Figura 14).

Ainda sobre o gráfico da figura 14, “brancos” e “amarelos” seguem um índice de óbito em proporções equivalentes, todavia, é observado que população autodenominada “parda” e principalmente “indígenas” possuem um índice de mortalidade numa curva muito inferior. Esta assimetria será investigada nos próximos capítulos.

Ao final desta análise, o dataset transformado foi salvo com o seguinte comando:

```
cor_se_melt.to_csv('mortalidade-cor-semana-tratado.csv')
```

Todo o código e saídas geradas na análise deste capítulo estão disponíveis no seguinte endereço:

https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/03%20-%20merge-covid-sem-pop.ipynb

4.2 Calculando o impacto das faixas etárias no número de óbitos

Apesar da análise do capítulo anterior revelar uma suposta vulnerabilidade social da população negra em detrimento dos demais grupos étnicos, esta análise ainda desconsidera que os grupos perfis etários diferentes, o que influencia na forma como as leituras devem ser feitas, já que a infecção por SARS-Cov-2 afeta mais, notadamente, pessoas de mais idade segundo o INSTITUTO POLIS(2019).

Portanto, este capítulo foca em análises focada nas faixas etárias desses grupos populacionais a fim de extrair mais insights e revelações sobre o real comportamento, impactos e assimetrias da pandemia nestes grupos.

Recapitulando a limpeza e tratamento de dados no “Dataset 2” (capítulo 3.3), que traz consigo os dados da população residente no município de São Paulo agrupados por faixa etária (de 5 em 5 anos) e cor/raça/etnia (vide imagem abaixo), serão aplicadas algumas transformações que permitirão análises mais profundas.

Cor	0 a 4	5 a 9	10 a 14	15 a 19	20 a 24	25 a 29	30 a 34	35 a 39	40 a 44	45 a 49	50 a 54
Branca	444168	433860	466843	462715	570950	636194	599762	524122	485640	465774	425601
Preta	29954	39476	52706	56272	68610	75869	73139	64755	59093	51910	47914
Amarela	8329	9020	10726	12276	17282	21097	20445	18141	17467	17002	17422
Parda	227650	275068	336289	309997	333423	339887	315436	280563	249819	207196	175970
Indígena	813	840	843	982	1352	1484	1256	1072	939	823	734

Figura 17: População residente em São Paulo agrupada por faixa etária e cor.

A primeira transformação será a transformação a fim de converter as colunas de faixas etárias em valores de linhas através do comando:

```
pop_melt = pop_norm.melt(id_vars=['Cor', 'Total'], value_vars=pop_norm.columns
[1:-1], var_name='Faixa Etária', value_name="Habitantes")
```

Depois, transformar os campos Habitantes (novo) e Total em valores numéricos com os comandos abaixo:

```
pop_melt.loc[:, 'Habitantes'] = pd.to_numeric(pop_melt['Habitantes'], errors='
coerce')
pop_melt.loc[:, 'Total'] = pd.to_numeric(pop_melt['Total'], errors='coerce')
```

Com isto, o dataset apresenta a seguinte estrutura (10 primeiras linhas):

	Cor	Total	Faixa Etária	Habitantes
0	Branca	6824668	0 a 4	444168
1	Preta	736083	0 a 4	29954
2	Amarela	246244	0 a 4	8329
3	Parda	3433218	0 a 4	227650
4	Indígena	12977	0 a 4	813
5	Branca	6824668	5 a 9	433860
6	Preta	736083	5 a 9	39476
7	Amarela	246244	5 a 9	9020
8	Parda	3433218	5 a 9	275068
9	Indígena	12977	5 a 9	840
10	Branca	6824668	10 a 14	466843

Figura 18: Dataset 2 transformado – etapa 1

Neste ponto é possível criar uma nova coluna para calcularmos a razão (% do total de habitantes) para cada linha através do comando abaixo:

```
pop_melt['Habitantes por total cor'] = pop_melt['Habitantes'] / pop_melt.loc[:
, 'Total']
```

O que resulta na estrutura abaixo (5 primeiras linhas):

	Cor	Total	Faixa Etária	Habitantes	Habitantes por total cor
0	Branca	6824668	0 a 4	444168	0.065083
1	Preta	736083	0 a 4	29954	0.040694
2	Amarela	246244	0 a 4	8329	0.033824
3	Parda	3433218	0 a 4	227650	0.066308
4	Indígena	12977	0 a 4	813	0.062649

Figura 19: Dataset 2 transformado – etapa 2

Com esta estrutura definida é possível visualizar graficamente a distribuição das faixas etárias dentro dos grupos étnicos.

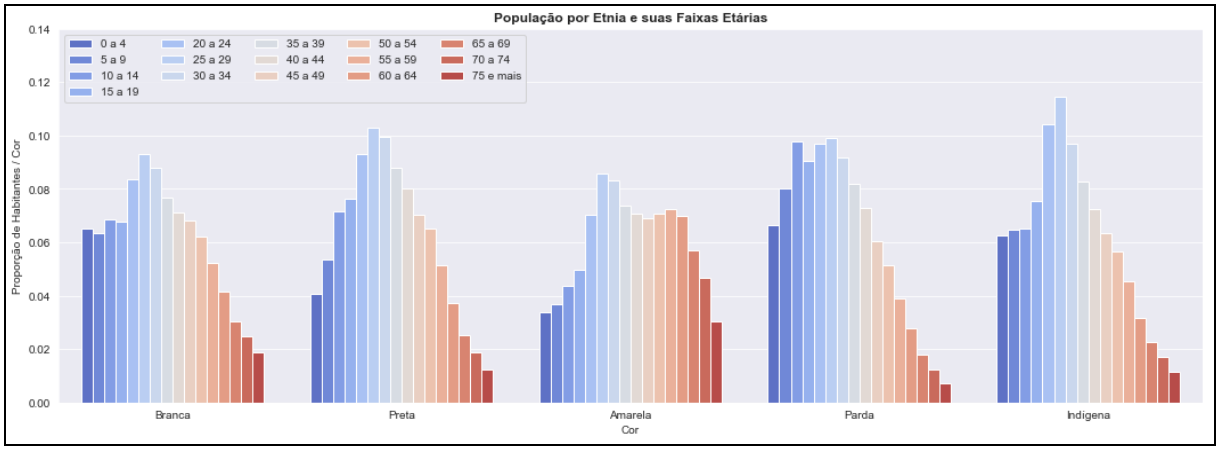

Gráfico 3: Concentração da população por Faixas Etárias e Etnias

Fica evidente no gráfico acima que a população branca e amarela concentra mais pessoas do grupo de risco – com 60 anos ou mais – e, portanto, sujeitas aos efeitos mais graves da COVID-19. No sentido inverso desta mesma análise, a taxa de mortalidade da população negra e parda pode mascarar o real impacto da infecção neste grupo, pelo fato de ser um grupo relativamente mais jovem, sugerindo de forma equivocada, que ele estaria mais protegido dos impactos da doença.

A nível de complemento de constatação, podemos comprovar a maior probabilidade de incidência de óbitos de acordo com maiores faixas etárias através de uma nova transformação, desta vez sobre o Dataset 3 (o qual traz o número de óbitos por cor e suas respectivas faixas etárias).

Desta vez o objetivo é de transformar as colunas de faixas etárias em linhas com seus respectivos números de óbitos através do comando abaixo:

```
cor_idade_melt = cor_idade.melt(id_vars=['Cor', 'Total'], value_vars=cor_idade
.columns[1:-2],
                                var_name='Faixa Etária', value_name='Óbitos')
```

Com isto, a estrutura do Dataset 3 passa a ser como na figura abaixo (10 primeiras linhas):

	Cor	Total	Faixa Etária	Óbitos
0	Branca	14297	0-4a	18.0
1	Preta	2108	0-4a	0.0
2	Amarela	544	0-4a	0.0
3	Parda	5367	0-4a	17.0
4	Indígena	16	0-4a	1.0
5	Total	23088	0-4a	38.0
6	Branca	14297	5-9a	8.0
7	Preta	2108	5-9a	1.0
8	Amarela	544	5-9a	0.0
9	Parda	5367	5-9a	5.0
10	Indígena	16	5-9a	0.0

Figura 20: Dataset 3 – Primeira transformação

Após uma série de transformações as quais podem ser observadas no código-fonte de referência (as quais envolvem criações de índices, loopings para o cálculo do número de habitantes por faixa etária, entre outros), é possível consolidar as informações na estrutura abaixo (5 primeiras linhas):

	Cor	Total	Faixa Etária	Óbitos	Habitantes	Óbitos por 100 mil
0	Branca	14297	0-4a	18.0	444168	4.052521
1	Preta	2108	0-4a	0.0	29954	0.000000
2	Amarela	544	0-4a	0.0	8329	0.000000
3	Parda	5367	0-4a	17.0	227650	7.467604
4	Indígena	16	0-4a	1.0	813	123.001230

Figura 21: Dataset 3 – última transformação

Isto permite duas análises gráficas acerca do índice de mortalidade em relação às faixas etárias – a primeira, onde observamos no gráfico abaixo que a taxa de óbitos aumenta drasticamente conforme a idade dos indivíduos:

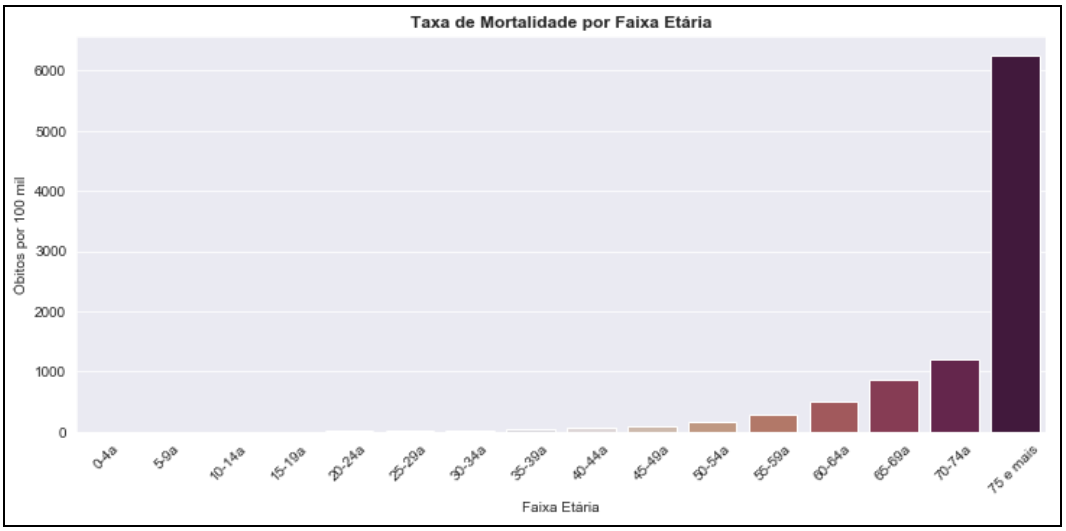

Gráfico 4: Taxa de Mortalidade por Faixa Etária

E a segunda, que ilustra que a baixa concentração de negros e pardos nas maiores faixas etárias, camufla a sua real taxa de mortalidade (que é sempre mais elevada proporcionalmente) em praticamente todas as faixas, como mostra o gráfico abaixo:

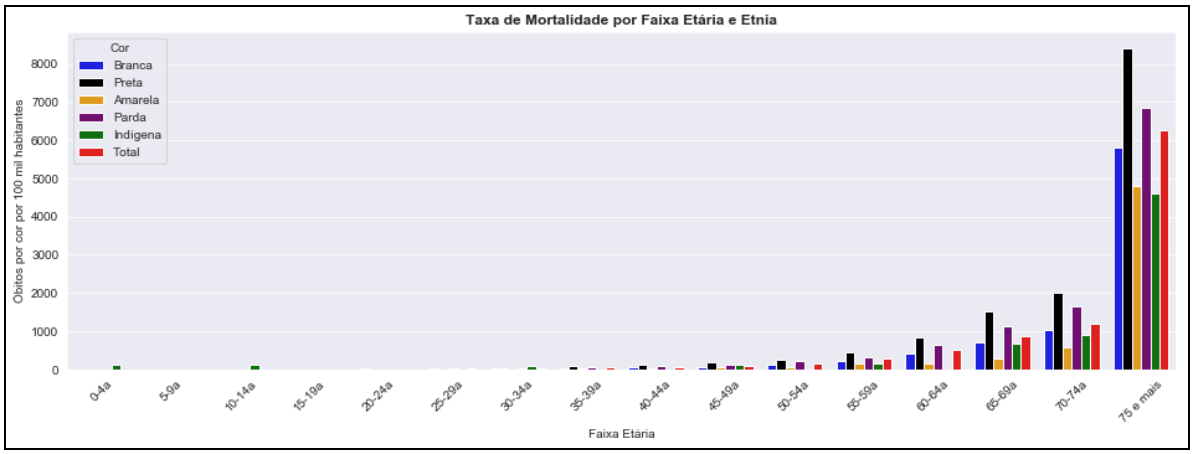

Gráfico 5: Taxa de Mortalidade por Faixa Etária e Etnia

Códigos e saídas disponíveis em:
https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/05%20-%20mortalidade-covid-idade.ipynb

5. Criação de Modelos de Machine Learning

Esta seção tem como objetivo encontrar um modelo matemático capaz de prever o número de mortes nos primeiros meses de 2021 com base nos dados coletados em 2020.

Como nas fontes de dados coletadas podem ser observadas informações registradas em intervalos iguais de tempos, será realizada uma análise com base em séries temporais.

Uma série temporal é uma sequência de observações sobre uma variável de interesse. A variável é observada em pontos temporais discretos, usualmente equidistantes, e a análise de tal comportamento temporal envolve a descrição do processo ou fenômeno que gera a sequência.

Em séries temporais a ordem dos dados é fundamental, os dados anteriores influenciam os dados futuros (auto-correlação), principalmente no caso deste trabalho onde algoritmos de machine learning tentarão prever a variável de interesse (número de óbitos) com base nos valores passados da mesma variável.

Embora existam muitas ferramentas que fornecem recursos de análise, visualização e desenvolvimento de modelos de machine learning, poucas fornecem recursos de construção de modelos de aprendizado profundo sem código de maneira extremamente rápida e intuitiva. Uma dessas ferramentas é o BigML (https://www.bigml.com). O BigML fornece aprendizado de máquina comoditizado como um serviço para analistas de negócios e integração de aplicativos. O objetivo do BigML é simples: para tornar o aprendizado de máquina fácil, simples e prático para todos os usuários. Um modelo de aprendizado de máquina ou aprendizado profundo (deep leaning) pode ser criado com apenas 3 a 4 cliques.

Antes de submeter o dataset 1 tratado para a plataforma, é interessante realizar algumas transformações para facilitar o cálculo com base em séries temporais.

A principal destas, é transformar o campo SEMANA (que possui apenas o número da semana no ano) para um formato de data (o qual seja entendido pelos algoritmos de machine learning como uma data válida). Para isso foram utilizadas bibliotecas como “time” e “datetime” para realizar as conversões em novas colunas.

O código completo das etapas de transformação estão disponíveis no arquivo: https://github.com/fredcobain/tcc_pos_data_pucmg/blob/master/dados_2021/07%20-%20ml-transform.ipynb

As transformações aplicadas resultam na estrutura da imagem abaixo (considerando as 10 primeiras linhas):

	Semana	Amarela	Branca	Indígena	Parda	Preta	Total	Semana_Num	Data	Data_Simp
0	SE_20 12	2.0	51.0	0.0	5.0	2.0	60.0	12	Mon Mar 23 00:00:00 2020	23/03/2020
1	SE_20 13	12.0	229.0	1.0	42.0	17.0	301.0	13	Mon Mar 30 00:00:00 2020	30/03/2020
2	SE_20 14	28.0	591.0	1.0	168.0	71.0	859.0	14	Mon Apr 6 00:00:00 2020	06/04/2020
3	SE_20 15	45.0	1028.0	3.0	339.0	135.0	1550.0	15	Mon Apr 13 00:00:00 2020	13/04/2020
4	SE_20 16	64.0	1470.0	3.0	526.0	190.0	2253.0	16	Mon Apr 20 00:00:00 2020	20/04/2020
5	SE_20 17	92.0	1971.0	4.0	733.0	268.0	3068.0	17	Mon Apr 27 00:00:00 2020	27/04/2020
6	SE_20 18	116.0	2605.0	4.0	1016.0	370.0	4111.0	18	Mon May 4 00:00:00 2020	04/05/2020
7	SE_20 19	146.0	3220.0	6.0	1317.0	456.0	5145.0	19	Mon May 11 00:00:00 2020	11/05/2020
8	SE_20 20	164.0	3905.0	7.0	1615.0	576.0	6267.0	20	Mon May 18 00:00:00 2020	18/05/2020
9	SE_20 21	183.0	4573.0	8.0	1940.0	661.0	7365.0	21	Mon May 25 00:00:00 2020	25/05/2020
10	SE_20 22	202.0	5260.0	8.0	2210.0	768.0	8448.0	22	Mon Jun 1 00:00:00 2020	01/06/2020

Figura 22: Dataset 1 transformado para ingestão na plataforma BIGML

Após ingestão do dataset transformado (conforme estrutura da Figura 22), a plataforma BigML testou centenas de algoritmos e modelos de séries temporais e trouxe 8 modelos de predição possíveis de acordo com a tabela abaixo:

Posição	Modelo	Descrição	AIC	AICc	BIC	R2
1	A,Ad,N	Método linear de tendência amortecido com erros aditivos. Erro aditivo, tendência ateída aditiva e nenhum modelo de sazonalidade.	529.43	531.9	539.71	0.9998
2	A,A,N	O método linear de Holt com erros aditivos. Erro aditivo, tendência aditiva e nenhum modelo de sazonalidade.	530.89	532.6	539.45	0.9998
3	M,Md,N	Método exponencial de tendências amortecido. Erro multiplicativo, tendência amortecido multiplicativa e nenhum modelo de sazonalidade.	672.41	674.88	682.69	0.9997
4	A,N,N	Simples suavização exponencial com erros aditivos. Erro aditivo, sem tendência e sem modelo de sazonalidade.	685.7	686.35	690.84	0.9918
5	M,A,N	O método linear de Holt com erros multiplicativos. Erro multiplicativo, tendência aditiva e nenhum modelo de sazonalidade.	710.33	712.04	718.9	0.9996
6	M,Ad,N	Método exponencial de tendência amortecido com erros multiplicativos. Erro multiplicativo, tendência ateída aditiva e nenhum modelo de sazonalidade.	719.92	722.39	730.2	0.9978
7	M,M,N	Método de tendência exponencial. Erro multiplicativo, tendência multiplicativa e nenhum modelo de sazonalidade.	722.88	724.59	731.44	0.9989
8	M,N,N	Simples suavização exponencial com erro multiplicativo. Erro multiplicativo, sem tendência e sem modelo de sazonalidade.	816.48	817.13	821.62	0.9646

Tabela 5: Algoritmos e modelos sugeridos pela plataforma BIGML

Dado o melhor modelo classificado pela plataforma (A, Ad, N), fazendo uma predição para o primeiro semestre de 2021, tem-se que o número de mortes acumuladas estarão próximos à casa dos 29 mil óbitos (vide gráfico abaixo):

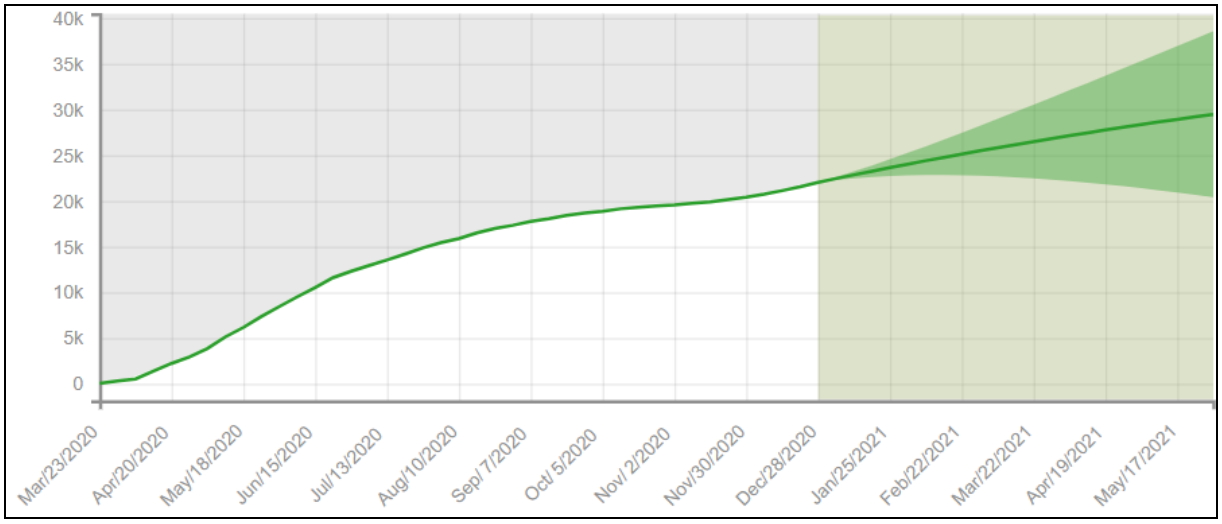

Gráfico 7: Predição utilizando o melhor modelo

Dentre os 3 modelos mais bem posicionados, o mais pessimista (M,M,N) sugere um número de óbitos próximo de 35 mil, enquanto o modelo mais otimista (M,Md,N) aponta para um plateau de controle tendendo a 24 mil óbitos para o primeiro semestre de 2021.

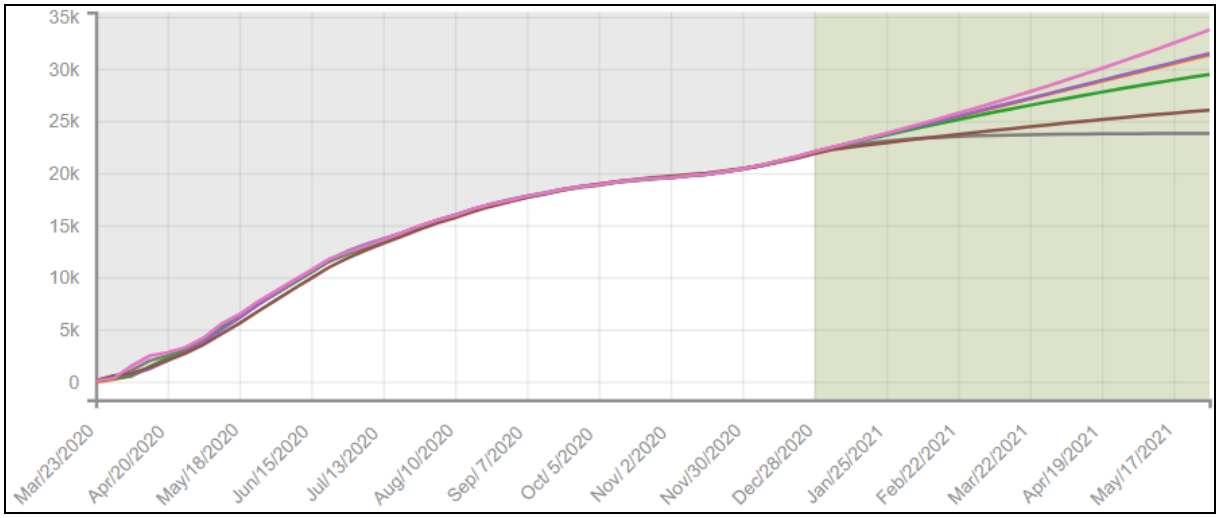

Gráfico 8: Comparativo entre os modelos de machine learning

6. Apresentação dos Resultados

Resumidamente, os seguintes passos foram realizados para atingir os objetivos deste trabalho foram:

a) Coleta de dados (realizada no capítulo 2)
b) Tratamento, Limpeza e Processamento de Dados (realizada no capítulo 3)
c) Análise e Exploração dos dados (realizada no capítulo 4)
d) Criação de Modelos de Machine Learning e Geração de Predições (realizada no capítulo 5)

As análises realizadas principalmente nos capítulos 4 e 5 trazem à tona as seguintes conclusões acerca dos dados coletados sobre óbitos de covid em São Paulo:

Embora a quantidade de óbitos da população branca em valores absolutos seja superior aos demais grupos, isto é apenas uma consequência direta do número de habitantes que pertencem a este grupo (aproximadamente 60% da população estudada).

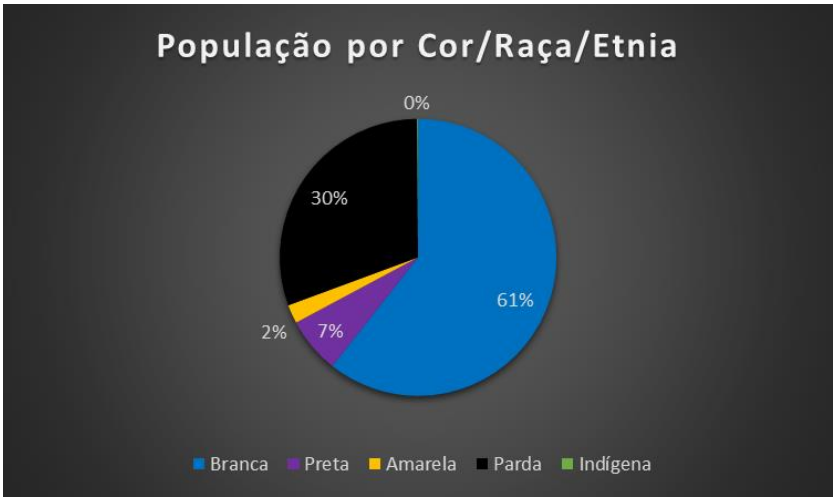

Gráfico 6: Distribuição da população estudada por cor/raça/etnia

A pandemia de COVID-19 realmente afeta a população mais idosa (conforme observado no Gráfico 4).

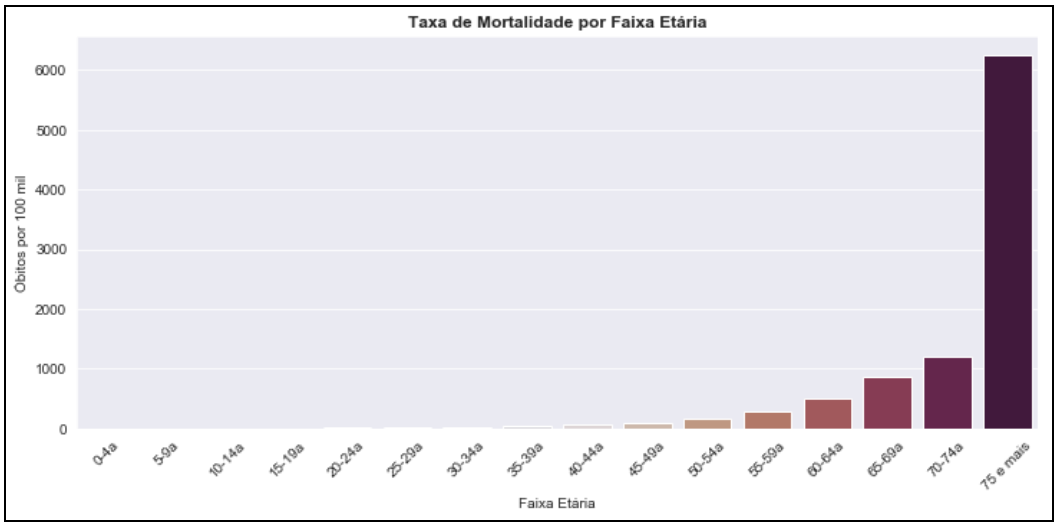

Gráfico 4: Taxa de Mortalidade por Faixa Etária

A curva de óbitos mais tênue em relação à população parda (Gráfico 2), deve-se principalmente à baixa concentração de idosos neste grupo.

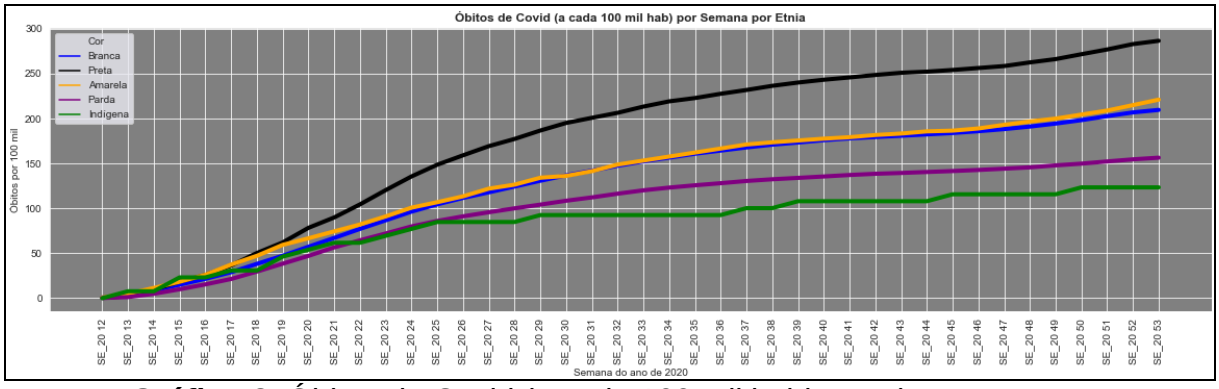

Gráfico 2: Óbitos de Covid (a cada 100 mil habitantes) por semana

A maior concentração de idosos está nos grupos de população amarela e branca (Gráfico 3).

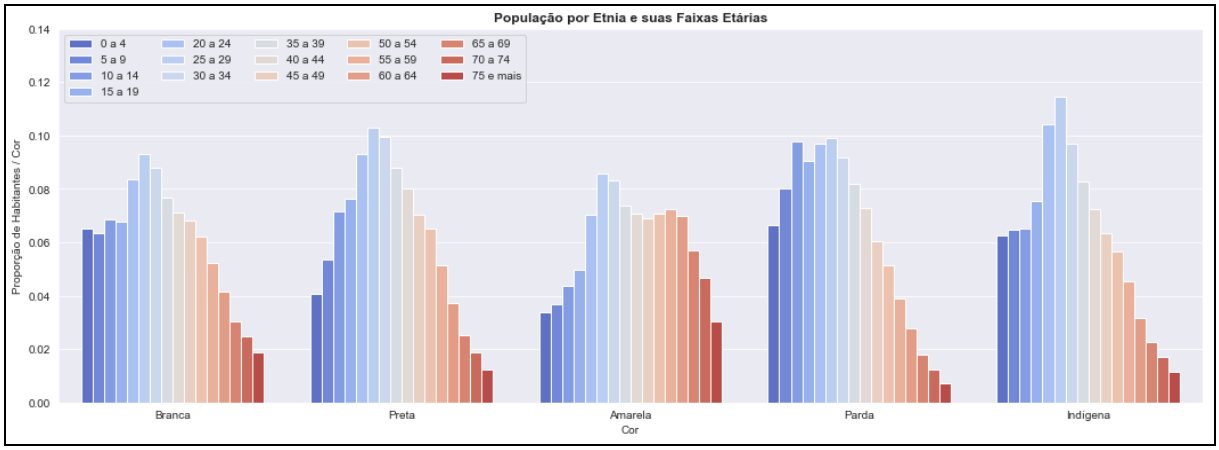

Gráfico 3: Concentração da população por Faixas Etárias e Etnias

Isso expõe ainda mais a vulnerabilidade da população negra, uma vez que esta tem uma curva de óbitos mais acentuada durante o tempo (Gráfico 2) ao mesmo tempo que possui uma baixa concentração de idosos (Gráfico 3) quando comparada à população amarela e branca.

Em todas as faixas etárias, quando aplicadas as corretas proporções populacionais, observa-se em todas as faixas etárias uma proporção de óbitos maior nas populações negras e pardas (Gráfico 5).

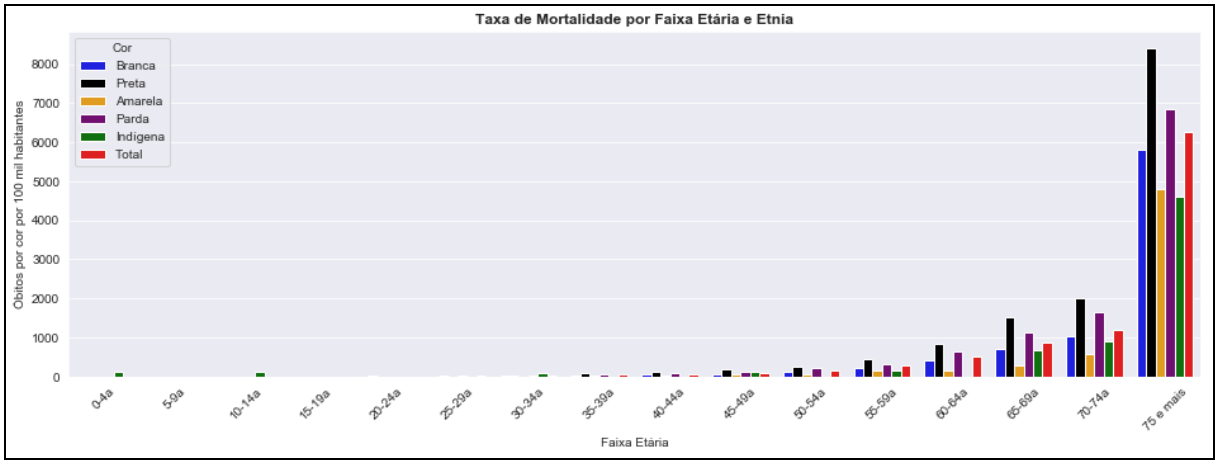

Gráfico 5: Taxa de Mortalidade por Faixa Etária e Etnia

Dado o melhor modelo classificado pela plataforma (A, Ad, N), fazendo uma predição para o primeiro semestre de 2021, tem-se que o número de mortes acumuladas estarão próximos à casa dos 900 mil óbitos (vide gráfico abaixo):

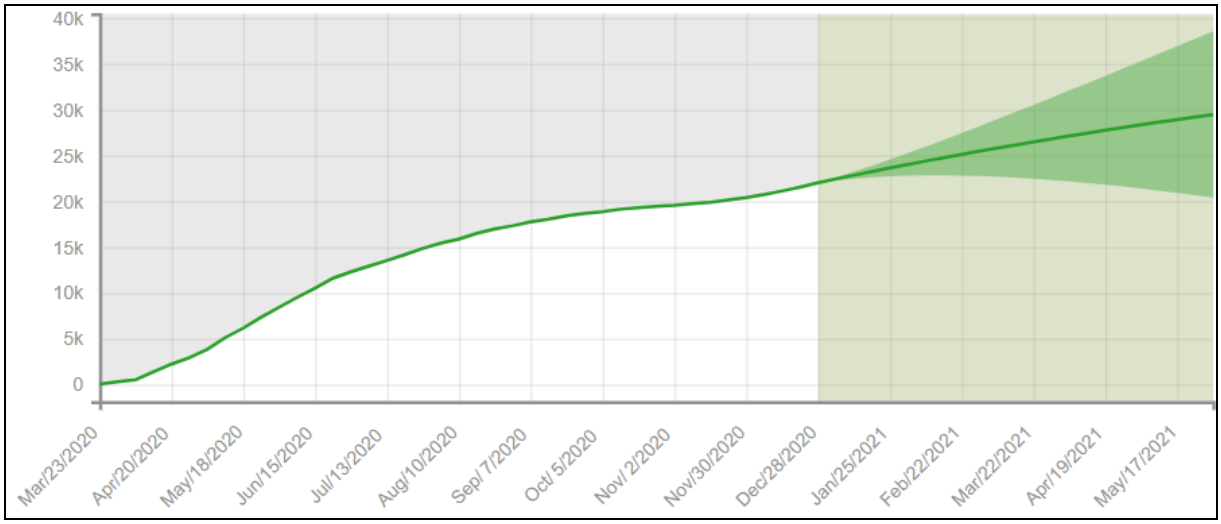

Gráfico 7: Predição utilizando o melhor modelo

Dentre os 6 modelos mais bem posicionados, o mais pessimista (A,A,N) sugere um número de óbitos acumulados acima de 1 milhão, enquanto o modelo mais otimista (M,Md,N) aponta para um plateau de controle tendendo a 660 mil óbitos para o primeiro semestre.

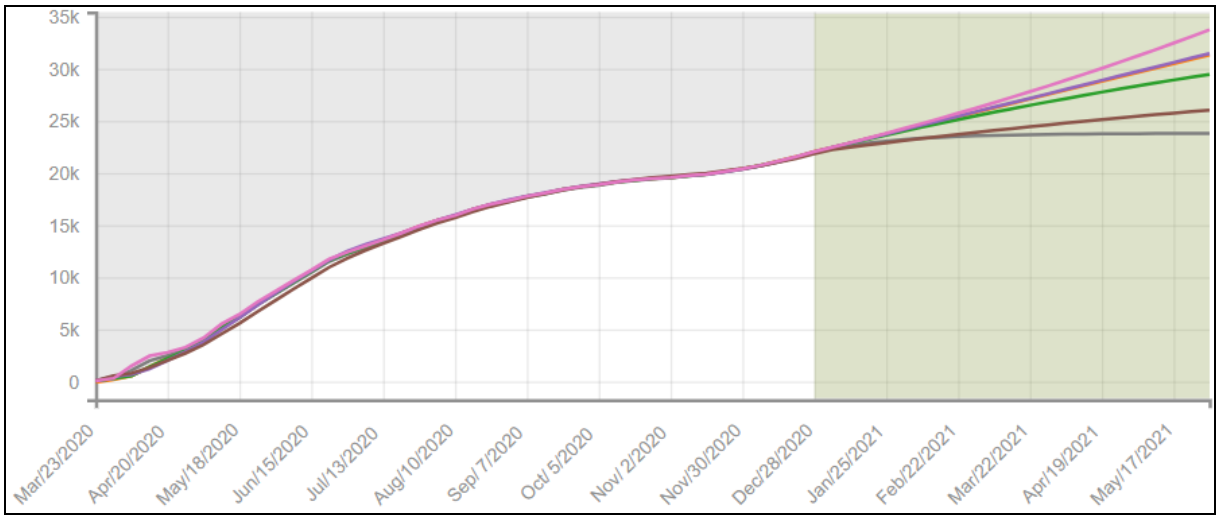

Gráfico 8: Comparativo entre os modelos de machine learning

O código fonte e saídas que proporcionaram os gráficos e análises acima estão disponíveis na URL abaixo:

https://github.com/fredcobain/tcc_pos_data_pucmg/tree/master/dados_2021

7. Links

Link para o repositório: https://github.com/fredcobain/tcc_pos_data_pucmg

Link para o vídeo: https://youtu.be/OF0f24SLQ5s

REFERÊNCIAS

Como a ciência de dados vem ajudando na luta contra a COVID-19. **CRITEO, 2020**. Disponível em: <https://www.criteo.com/br/blog/como-a-ciencia-de-dados-vem-ajudando-na-luta-contra-o-covid-19/>. Acesso em: 01 de jul. 2020

O que é ciência de dados. **ORACLE, 2020**. Disponível em : <https://www.oracle.com/br/data-science/what-is-data-science>. Acesso em: 01 de jul. 2020

O que é ciência de dados e como aplica-la nos negócios. **AQUARELA, 2018**. Disponível em : < https://www.aquare.la/o-que-e-ciencia-de-dados-data-science-para-negocios/>. Acesso em: 02 de jul. 2020

Data scientist, the sexiest job of the 21st century. **HBR**. Disponível em: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>. Acesso em: 01 de mar. 2020

Sistema TABNET SP - DATASUS. **TABNET**. Disponível em: <http://tabnet.saude.prefeitura.sp.gov.br>. Acesso em: 04 de jul. 2020

Instituto Polis - Raça e COVID no município de São Paulo. **POLIS**. Disponível em: <https://polis.org.br/estudos/raca-e-covid-no-msp/>. Acesso em: 20 de jun. 2020

ALMEIDA, S. **Racismo Estrutural**. São Paulo: Editora Pólen Livros, 2019.

SCIELO. População negra e Covid-19: reflexões sobre racismo e saúde. **SCIELO** Disponível em: <https://www.scielo.br/j/ea/a/LnkzjXxJSJFbY9LFH3WMQHv/?lang=pt>. Acesso em: 02 de jul. 2020

ATELIWARE. Feature Engineering: Preparando dados para o aprendizado de máquina. **ATELIWARE**. Disponível em: <https://ateliware.com/blog/feature-engineering>.. Acesso em: 04 de jul. 2020

IBGE. Desigualdades sociais por cor ou raça, 2019. **IBGE**. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/25844-desigualdades-sociais-por-cor-ou-raca.html?=&t=sobre>. Acesso em: 01 de jul. 2020

PYTHON. Disponível em: <https://www.python.org/>. Acesso em: 04 de jul. 2020

JUPYTER. Disponível em: <https://jupyter.org/>. Acesso em: 04 de jul. 2020

ANACONDA. Disponível em: <https://www.anaconda.com>. Acesso em: 04 de jul. 2020