Projeto de Análise Exploratória e Visualização de Dados da Covid-19 em Manaus

Adham Lucas da Silva Oliveira[1], **Alex T. Carvalho**[1], **Enrique L. B. Izel**[1], **Nayara da Silva Cerdeira da Costa**[1], **Vitor Simões Azevedo**[1]

[1]Núcleo de Computação – Universidade do Estado do Amazonas (UEA)
Manaus – AM – Brasil

A análise deste trabalho está disponível em: github.com/EnriqueIzel2/top-analise-redes-neurais

Abstract. *This work describes the obtained results in an analysis of the COVID-19 confirmed cases data in the city of Manaus. With cleaning steps, dataset organization and with didatic's goals.*

Resumo. *Este trabalho descreve os resultados obtidos em uma análise exploratória e visualização dos dados de casos confirmados da COVID-19 na cidade de Manaus. Com etapas de limpeza e organização do dataset com objetivos didáticos.*

1. Introdução

A Covid-19 teve seu inicio no fim de 2019 na cidade de Wuhan na China, com sua alta taxa de disseminação geográfica a Organização Mundial de Saúde (OMS) declarou em março de 2020 a pandemia pelo novo coronavírus (SARS-COV-2). [ORELLANA 2020] No Brasil com o primeiro caso no dia 26 de fevereiro, os dados oficiais de óbitos pelo coronavírus, quantidade de infecções e curados são disponibilizados diariamente pelo Ministério da Saúde. [Vanessa Aquino 2020] O Amazonas foi o primeiro Estado do Brasil a se tornar o epicentro da pandemia da Covid-19 em abril de 2020. [ORELLANA 2020] Este relatório apresenta uma análise exploratória do cenário da Covid - 19 em Manaus, foi utilizado nesse projeto os dados disponibilizados pela Prefeitura de Manaus.

2. Metodologia

Para esse trabalho foi adotada a seguinte metodologia:

a) Coleta de dados:
   Extração dos dados no dia 06 de agosto de 2020
b) Tratamento de Dados:
   Desenvolvimento de scripts em Python, utilizado as bibliotecas Pandas e Numpy para tratamento de dados.
c) Visualização de Dados:
   Desenvolvimento de scripts em Python, utilizando as bibliotecas matplotlib, seaborn e plotly.

3. Tratamento de Dados

O Dataset utilizado foi disponibilizado pela Prefeitura de Manaus e inicialmente era composto por 108351 instâncias, com os atributos listados na Figura 1.

Após processamento para considerar apenas casos confirmados e ignorar atributos relativos à comorbidades, etnia, sintomas, outras datas que não fosse data de notificação, profissão, origem, outros atributos, também foram desconsiderados atributos com campos nulos, porém no atributo ”tipo teste”os campos nulos foram tratados como ”Desconhecido”. Isso em decorrência da grande quantidade de campos nulos no atributo tipo teste que após a exclusão dos campos nulos reduziria drasticamente o número de dados disponíveis para análise e o tratamento da data de notificação considerando somente casos ápos a data do primeiro caso confirmado pelo Estado do Amazonas. Este processamento culminou na estruturação de um dataset com 12658 dados com os atributos apresentados na Figura 2.

#	Column	Non-Null Count	Dtype
0	_idade	108230 non-null	float64
1	_faixa etária	108351 non-null	object
2	_sexo	107551 non-null	object
3	_bairro	106223 non-null	object
4	_classificacao	108351 non-null	object
5	_comorb_renal	94992 non-null	object
6	_comorb_diabetes	95646 non-null	object
7	_comorb_imuno	94859 non-null	object
8	_comorb_cardio	95698 non-null	object
9	_conclusao	64812 non-null	object
10	_dt_notificacao	108340 non-null	object
11	_taxa	102551 non-null	object
12	_dt_evolucao	39685 non-null	object
13	_raca	83920 non-null	object
14	_dt_sintomas	104710 non-null	object
15	_criterio	8182 non-null	object
16	_tipo_teste	72500 non-null	object
17	_sintoma_garganta	98279 non-null	object
18	_sintoma_dispneia	99064 non-null	object
19	_sintoma_febre	99189 non-null	object
20	_sintoma_tosse	99132 non-null	object
21	_sintoma_outros	99133 non-null	object
22	_etnia	35 non-null	object
23	_profiss_saude	90083 non-null	object
24	_srag	9016 non-null	object
25	_se_notificacao	108340 non-null	float64
26	_distrito	102551 non-null	object
27	_bairro_mapa	102551 non-null	object
28	_comorb_respiratoria	95003 non-null	object
29	_comorb_cromossomica	94890 non-null	object
30	_comorb_hepatica	3296 non-null	object
31	_comorb_neurologica	3373 non-null	object
32	_comorb_hemato	3341 non-null	object
33	_comorb_obessidade	3250 non-null	object
34	_origem	108351 non-null	object
35	_evolução	12462 non-null	object

Figura 1. Atributos do dataset disponibilizado pela Prefeitura de Manaus

#	Column	Non-Null Count	Dtype
0	_idade	12658 non-null	int64
1	_faixa etária	12658 non-null	object
2	_sexo	12658 non-null	object
3	_classificacao	12658 non-null	object
4	_conclusao	12658 non-null	object
5	_dt_notificacao	12658 non-null	datetime64[ns]
6	_tipo_teste	12658 non-null	object
7	_distrito	12658 non-null	object
8	_bairro_mapa	12658 non-null	object

Figura 2. Atributos do dataset após limpeza

4. Análise Exploratória dos casos confirmados

Com o dataset contendo somente casos confirmados até a data de extração dos dados foi possível realizar algumas análises para responder algumas questões:

Apesar da alta transmissibilidade da Covid-19 a porcentagem de recuperados foi de 83.92%. Dos casos confirmados os que mais foram acometidos foram indivíduos do sexo masculinos com uma diferença de 0,5372%. Também foi possível notar que a media de idade é de 47.6 anos, e o desvio padrão de 18 anos, o *dataset* indica que a pessoa mais jovem a contrair o Covid-19 possui a idade de 0 anos, isso se deve provavelmente, a crianças que ainda não possuem 1 ano completo. E a pessoa mais velha na idade de 99 anos. Vale ressaltar que foram considerados idade até 100 anos, por julgar que dados maiores que 100 poderiam se tornar ruidosos para a análise.

No âmbito geográfico, o bairro que apresentou maior incidência de casos foi o CIDADE NOVA com 858 casos confirmados dos quais 719 se recuperaram, sendo o bairro com maior número de recuperados seguido pelos bairros de FLORES e ALVORADA, com 437 e 430 recuperados respectivamente.

Para que sejam tomadas decisões certas em favor da saúde pública, primeiro é necessário identificar bem a magnitude da ameaça à população, hoje, o novo coronavírus (SARS-CoV-2, causador da Covid-19). Isto é possível por meio dos testes para detectar a doença e, por essa razão, a testagem no maior número possível de cidadãos é fundamental para enfrentar o vírus, conforme recomendação da Organização Mundial da Saúde (OMS). [(IFF/Fiocruz) 2020] Os testes realizados em Manaus, segundo o *dataset* estão na Tabela 1.

	quantitativo	porcentagem
_tipo_teste		
DESCONHECIDO	6450	50.955917
ECLIA IgG	4	0.031601
ELISA IgM	6	0.047401
RT-PCR	1497	11.826513
TESTE RÁPIDO - ANTICORPO	3601	28.448412
TESTE RÁPIDO - ANTÍGENO	1100	8.690156

Tabela 1. Testes realizados em Manaus.

A taxa de letalidade é a proporção entre o número de mortes por uma doença e o número total de doentes que sofrem dessa doença, ao longo de um determinado período de tempo. Em Manaus a Covid-19 teve uma taxa de letalidade de 16.07% no período de 16 de Abril de 2020 até o dia 31 de Julho de 2020. Sendo 16 de Abril o primeiro caso confirmado de Covid-19 em Manaus, foi feita a limpeza de alguns casos com data de notificação anterior ao primeiro caso.

A fim de saber a correlação entre a idade e os casos de Covid-19, calculamos a correlação de Pearson com esses dois atributos e obtivemos o resultado de -0.08 o que nos dá o entendimento de uma correlação desprezível, ou seja, a idade de alguém não está relacionada com sua confirmação para Covid-19. [Mukaka 2012]

5. Visualização de Dados

Para a plotagem dos gráficos neste relatório foram utilizadas as bibliotecas *Matplotlib*, feita para o Python e sua extensão de matemática NumPy, *Seaborn*, uma biblioteca baseada no Matplotlib, mas com gráficos mais atrativos, e *Plotly*, que é uma biblioteca própria, mas com o mesmo propósito do *Seaborn*.

A respeito da distribuição dos casos geograficamente em Manaus destacamos os 10 bairros com maior incidência de casos, na Figura 3.

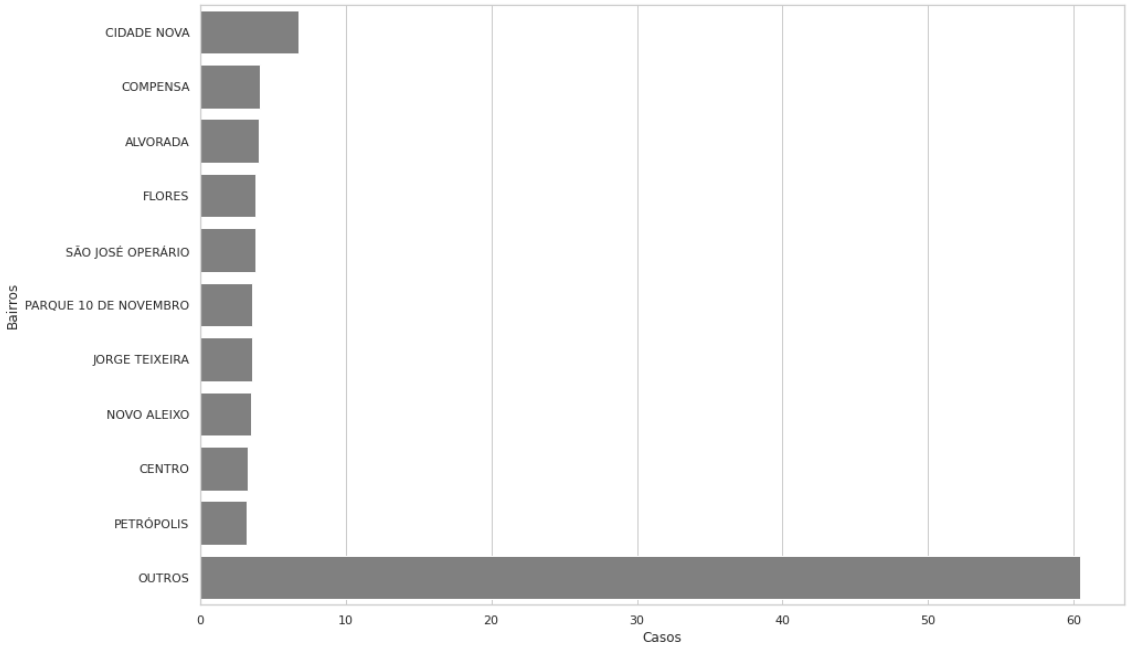

Figura 3. Os 10 bairros com maior incidência de casos

É possível notar que o bairro Nova Cidade tem um percentual maior de casos enquanto os demais bairros seguem em uma percentual semelhante. Já quando denotamos os casos confirmados por sexo, por idade percebemos que não há *outliers* o que sugere o coronavírus é contagioso em todas as faixas etárias, como mostrado no *boxplot* da Figura 4. Além de que o vírus não se mostra com maior número de casos em função de determinado sexo.

Visualizando por faixa etária, na Figura 5 percebemos que as faixas etárias que mais testaram positivo para covid-19 foram as pessoas que estão em idade ativa no mercado de trabalho de 20 a 59 anos, o que pode ser atribuído às pessoas que não estavam em isolamento devido o trabalho.

Levando em consideração o alto contágio, é importante analisar a evolução dos novos casos por dia. No gráfico da Figura 6 apresentamos os últimos 10 dias contidos no *dataset*.

Geralmente nas segundas-feiras há um salto no número de casos pois os casos do fim de semana só são registrados na segunda-feira, uma exceção observada sendo o dia 24 de julho, que foi uma sexta-feira.

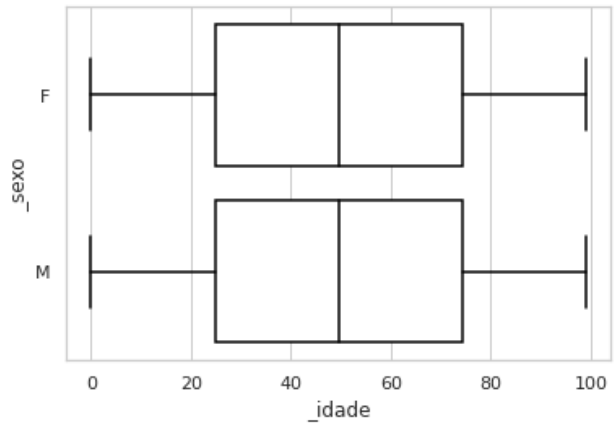

Figura 4. *Boxplot* da idade dos casos confirmados.

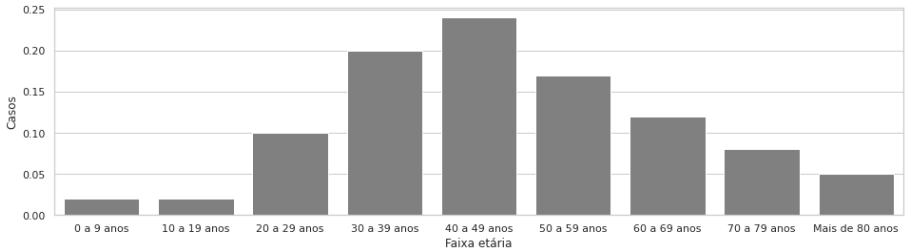

Figura 5. Faixa etária de cada caso confirmado.

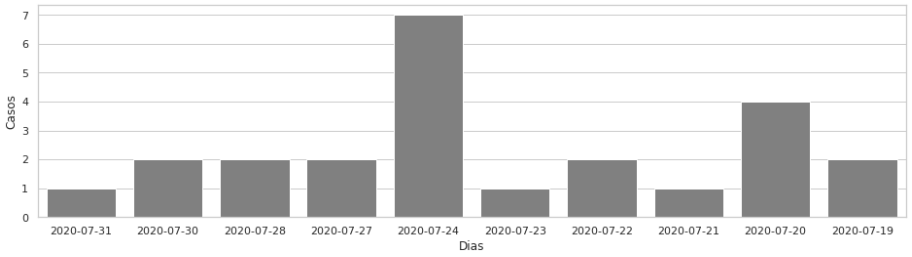

Figura 6. Novos casos nos últimos 10 dias contidos no *dataset*.

Os indivíduos que estão recuperados são aqueles que tiveram dois resultados negativos com pelo menos um dia de intervalo. Já nos casos leves de COVID-19, a OMS estima que o tempo entre o início da infecção e a recuperação dure até 14 dias. No gráfico de recuperados nos últimos 10 dias contido no *dataset*, na Figura 7, notamos que o número de recuperados é superior aos novos casos.

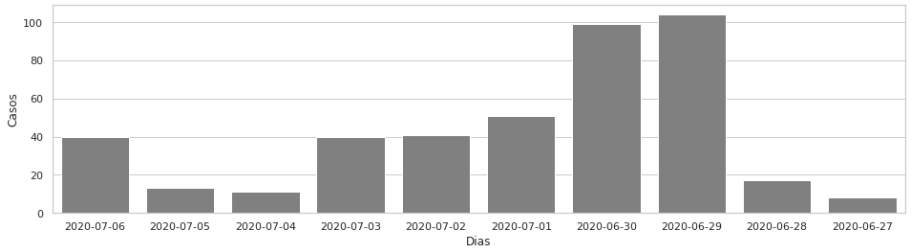

Figura 7. Casos recuperados nos últimos 10 dias contidos no *dataset*.

Sobre a curva de casos notificados, notamos uma crescente na curva no mês de maio, onde tivemos a superlotação [UOL 2020], e em meados do mês de junho notamos a diminuição da inclinação da curva de novos casos, como mostrado na Figura 8.

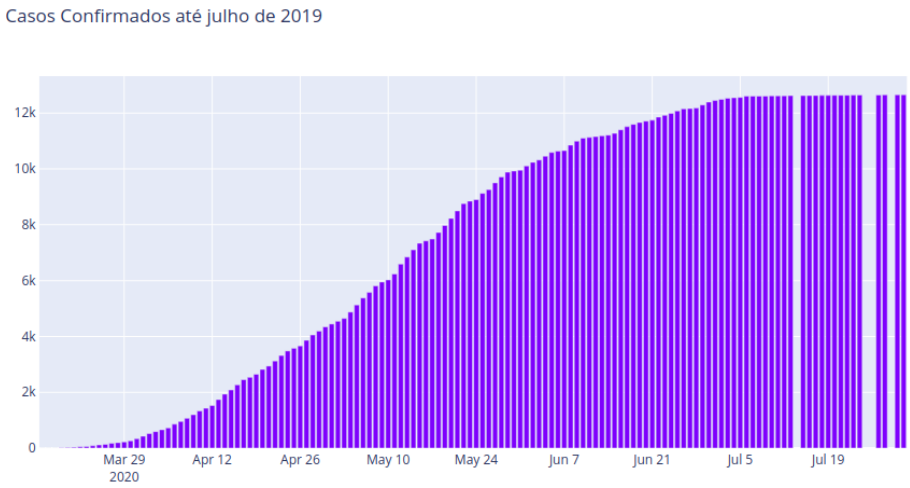

Figura 8. Casos recuperados nos últimos 10 dias contidos no *dataset*.

Há colunas em branco pois foram retirados dados que não estavam completos. E por fim, um gráfico do tipo *scatterplot*, denotando a idade relacionada ao número total de casos registrados para aquela idade. Pelo gráfico, mostrado na figura 9, existe uma tendência entre pessoas de 40 anos em serem infectadas, pessoas estas que provavelmente estão entre os que mais trabalham e que durante o período de quarentena, devido ao modelo de trabalho, tinham de sair para conseguir renda à família.

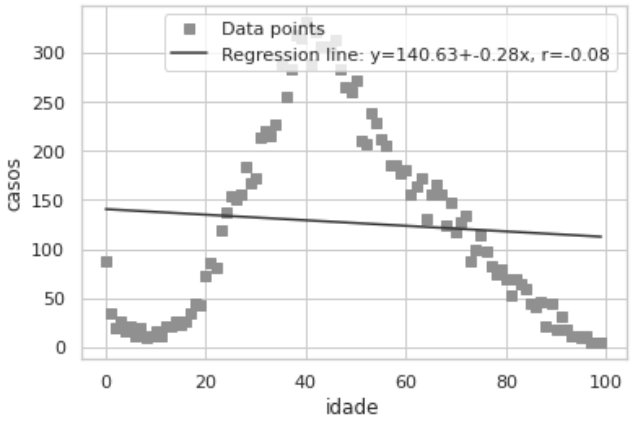

Figura 9. Scatterplot de idade x casos confirmados, mais correlação de Pearson.

6. Tipos de Tarefas

Um sistema de aprendizado é um programa de computador que toma decisões baseado em experiências acumuladas através da solução bem sucedida de problemas anteriores. Existem dois tipos de Aprendizado Supervisionado:

1. **Uma tarefa de classificação mediante Aprendizado Supervisionado que poderia ser feita com esta base de dados. Qual seria o atributo-alvo? Quais**

métricas de desempenho poderiam ser aplicadas? Que tipo de validação seria apropriado?
O atributo seria "Classificação". Ou seja, ele tendo o COVID-19 ou não.
Como se trata de algo da área de saúde, a especificidade (quantos negativos foram acertados) garante uma melhor qualidade para o modelo, afinal, estar doente e receber o diagnóstico de negativo é muito mais grave do que o contrário. Precisão é outra métrica importante, afinal, o modelo também precisa ser bom em detectar quem está infectado.

2. **Uma tarefa de regressão mediante Aprendizado Supervisionado que poderia ser feita com esta base de dados. Qual seria o atributo-alvo? Quais atributos preditores a equipe considera relevantes para o cenário?**
Uma tarefa de regressão, com esta base de dados, seria inconcebível, pois regressão, por definição retorna um número, e o único atributo que retornam um número é a idade, atributo que não convêm muito fazer alguma predição, ainda mais já existindo o atributo "faixa etária"
3. **Bônus: Qual tarefa de Aprendizado Não-Supervisionado poderia ser concebida neste contexto?**
Clusterização, para ver a relação entre idade e casos confirmados, sexo e casos confirmados, e a partir daí tentar clusterizar os dados e tentar enxergar alguma correlação.

Referências

(IFF/Fiocruz), M. M. M. (2020). Testes para a covid-19: como são e quando devem ser feitos. https://portal.fiocruz.br/noticia/testes-para-covid-19-como-sao-e-quando-devem-ser-feitos. Acessado em: 17 de agosto de 2020.

Mukaka, M. (2012). Statistics corner: A guide to appropriate uso of correlation coefficient in medical research.

ORELLANA, J. D. Y. e. a. (2020). Explode mortalidade em manaus, epicentro amazônico da covid-19. http://cadernos.ensp.fiocruz.br/csp/artigo/1101/explosao-da-mortalidade-no-epicentro-amazonico-da-ep. Acessado em: 17 de agosto de 2020.

UOL (2020). Covid-19: apesar de superlotado, hospital público de manaus tem andar vazio. https://noticias.uol.com.br/saude/ultimas-noticias/redacao/2020/05/03/covid-19-apesar-de-superlotado-hospital-publico-de-manaus-tem-andar-v
htm. Acessado em: 17 de agosto de 2020.

Vanessa Aquino, N. M. (2020). Brasil confirma primeiro caso da doença. https://www.saude.gov.br/noticias/agencia-saude/46435-brasil-confirma-primeiro-caso-de-novo-coronavirus. Acessado em: 17 de agosto de 2020.