Análise exploratória de um conjunto de dados sobre a COVID-19

Amanda L. M. Chaves.[1], **Erik A. S. Rey**[1], **Fílip A. A. Silva.** [1], **Nadine C. Brito**[1], **Wilbert L. E. Marins**[1]

[1]Núcleo de Computação – Universidade do Estado do Amazonas (UEA)
Caixa Postal 15.064 – CEP: 69050-020 – Manaus – AM – Brazil

{almc.ads,easr.snf17,faads.snf,ndcb.snf16,wlem.snf17}@uea.edu.br

Abstract. *The crisis situation in the world public health in 2020 due to the pandemic of COVID-19, creates the population's interest in the data collected in this atypical condition and the treatment and analysis of data is indispensable for a reliable overview of the scenario. In order to generate this overview, an analysis of a set of data on COVID-19 in the city of Manaus was carried out, and will be presented in this article.*

Resumo. *A situação de crise na saúde pública mundial em 2020 por conta da pandemia de COVID-19, cria o interesse da população nos dados levantados nessa condição atípica e o tratamento e a análise desses dados é indispensável para uma visão geral confiável do cenário. Com o intuito de gerar essa visão geral, a análise de um conjunto de dados sobre a COVID-19 na cidade de Manaus foi realizada, e será apresentada nesse artigo.*

1. Metodologia Utilizada

A primeira etapa para o desenvolvimento de tal pesquisa é a preparação do ambiente de desenvolvimento que será utilizado. Neste caso foi utilizada a linguagem de programação Python 3.6+ juntamente com as bibliotecas pandas e numpy para o primeiro Jupyter Notebook e a biblioteca matplotlib para o segundo. Para manter a pesquisa realizada em conjunto organizada, foi necessária a criação de um repositório `https://github.com/erikatilio/RNA-PP1.git` na plataforma GitHub, para hospedagem de código-fonte com controle de versão, a fim de facilitar o desenvolvimento e manutenção do código.

A análise foi baseada em um conjunto de dados disponibilizado pela Prefeitura da cidade de Manaus (`https://covid19.manaus.am.gov.br/wp-content/uploads/Manaus.csv`), até o dia 05/08/2020 que foi a data ao qual foi realizado o download da base de dados. Após a aquisição dos dados, através da linguagem de programação escolhida juntamente com as bibliotecas, foi necessária a aplicação de uma filtragem nos atributos considerados irrelevantes e exclusão das linhas com dados faltantes aos campos que serão utilizados. Em seguida foram respondidas algumas questões propostas pela impulsora deste trabalho, tais respostas, assim como as questões e as conclusões observadas pelos autores deste artigo serão discorridos a seguir.

2. Visão Geral dos Casos Confirmados

Foi feito uma limpeza nos dados, para a análise ser mais precisa, considerando o contexto do problema. Onde foram excluídos atributos, que não eram relevantes no contexto

de análise de casos de Covid-19 em Manaus. Atributos como tipo de comorbidades, sintomas, etnia, profissão, outras datas que não eram de notificação, origem e outros atributos que não estavam envolvidos no contexto do trabalho.

Todo e qualquer dado resultante abaixo leva em consideração a simplificação citada acima.

Os atributos considerados para análises posteriores resultam em 12, sendo eles: idade, faixa etária, sexo, bairro, classificação (no caso só os confirmados), conclusão (se o paciente se recuperou ou não), data de notificação, taxa, tipo de teste realizado , bairro mapa e distrito.

Até a data de 05 de agosto de 2020 em Manaus já foram confirmados 6145 casos confirmados com o vírus. Desses, 99.79% já encontram-se recuperados da doença, equivalendo a um total de 6132 pacientes.

Levando em consideração as datas de notificação informadas dentro da filtragem realizada na base de dados, o primeiro registro de caso confirmado de COVID-19 é do dia 30 de janeiro de 2020 e o último, até então, é do dia 05 de agosto de 2020.

Os casos acometeram mais indivíduos do sexo feminino com 3463 casos.

A partir da análise dos dados, pôde-se perceber que média da idade dos casos confirmados com a doença é de 41.49 anos (fase adulta) na faixa etária de 35 a 55 anos, sendo que o mais jovem possui menos de um ano de idade, e o mais idoso 99 anos de idade.

Ao fazer uma análise de incidência dos casos levando em consideração a localização, o barro da Cidade Nova segue como maior bairro com casos confirmados, com um total de 296 casos. Felizmente, houve um percentual de 100% casos recuperados dentre os confirmados na Cidade Nova. Os bairros que seguem com maiores casos recuperado dentro da cidade sao, respectivamente: Flores com 262 casos recuperados e Tarumã com 225 casos recuperados.

Ao total, foram realizados 5 testes nos pacientes com casos confirmados, sendo eles: ECLIA IgG são as células de memória, reagem positivo caso o paciente possuia memória para combater o vírus, o ELISA IgM que indica se o vírus está ativo no momento atual, ambos são marcadores no corpo que indicam se o paciente teve contato com o virus; o teste rápido - Antígeno reage positivo quando o paciente já esteve em contato como o virus, o teste rápido - Anticorpos que reage ao antígeno e o RT-PCR que reconhece o vírus caso ele esteja presente na amostra analisada. O teste mais realizado foi o de Anticorpos com um total de 3553, seguido pelo RT-PCR com 1486 teste e pelo teste rápido de Antígeno com um total de 1099.

Na cidade de Manaus a taxa de letalidade da doença foi equivalente a 16.17% dentre um quantitativo total de 12586 casos,isso se deve ao interesse em verificar as confirmações de infecção nos indivíduos que não efeturam nenhum tipo dos testes disponíveis. Taxa indicativa de considerações de casos baseados em atributos diferentes no ponto de vista médico.

Foi calculado o coeficiente da correlação de Pearson entre os atributos: idade e número de casos que foi igual a -0.223. Indicando assim, que a natureza dessa correlação

é negativa e de baixa intensidade.

3. Visualização dos Dados

A seção de visualização, é uma parte bastante intuitiva em uma análise de dados por conta do seu apelo visual e deixa mais fácil o entendimento de representações quantitativas dos mais diversos cenários. Assim sendo, foi utilizado desse método para apresentar algumas relações entre dados interessantes para esse momento.

Casos por bairro: No seguinte histograma, temos explicitado no eixo x, os 10 bairros com maior quantidade de casos confirmados, seguidos por uma categoria chamada ”Outros” que engloba todos os outros bairros da cidade de Manaus e no eixo y estão marcados as quantidade de casos de maneira percentual. Nesta tarefa, foi feito uma filtragem, com os 10 bairros e suas respectivas quantidades de casos e depois foi feito a mesma filtragem com os bairros restantes, e foi feito o somatório de suas quantidades de casos. E depois foi calculado a porcetagem de cada valor. E depois, essas informações foram utilizadas na plotagem do gráfico.

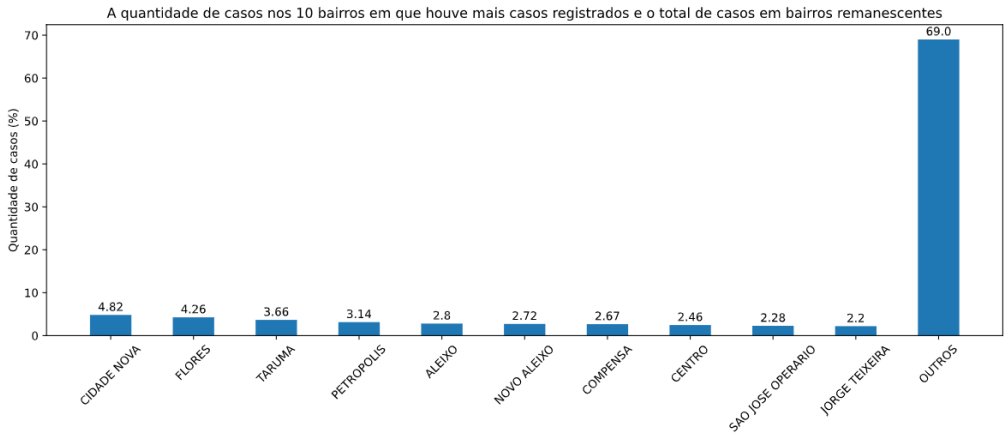

Figure 1. Quantidade de casos por bairro.

Comparativo entre sexos: Nos seguintes boxplots é evidenciada a diferença de casos confirmados entre homens e mulheres (eixo x) usando como métrica a idade dos individuos (eixo y). A existência de outliers traz a reflexão que este vírus afeta várias faixas etárias e ambos os sexos, desde crianças com menos de 1 ano de vida até pessoas a beira do centenário.

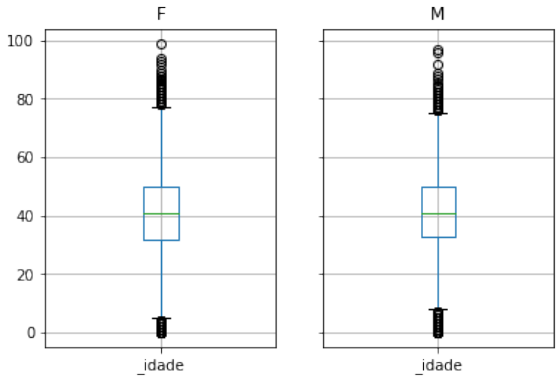

Figure 2. Relação idade por sexo.

Novos Casos por dia: No gráfico seguinte temos os dados dos 10 últimos dias da base de dados, pode-se observar que não são datas sequenciais, pois foi usado somente os dados tratados e algumas datas foram retiradas pois suas tabelas tinha dados faltantes. No eixo x, temos as datas dos 10 últimos dias registrados e no eixo y temos a quantidade de dados reportados. Nesta tarefa de visualização, foi feito a filtragem, levando em consideração o atributo de data de notificação, que é referente as datas, foi utilizado um array de datas, para armazenar as respectivas datas, e foi usado uma função chamada autolabel, para mostrar a quantidade de casos na parte de cima das barras.

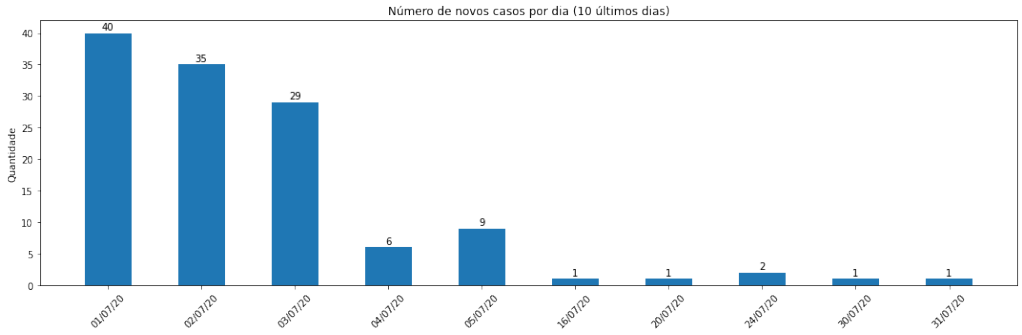

Figure 3. Novos casos por dia.

Casos recuperados: A mesma idéia do gráfico anterior foi aplicado neste gráfico, onde foi pego as informações dos ultimos 10 dias mas agora lidando com a informação dos casos onde os pacientes se recuperanram da doença. No eixo x podemos ver as respectivas 10 últimas datas, e no eixo y temos a quantidade de casos recuperados. Nesta tarefa, algumas coisas da questão anterior foram reutilizadas, só a filtragem mudou, que era o atributo conclusão, onde foram usados só os recuperados.

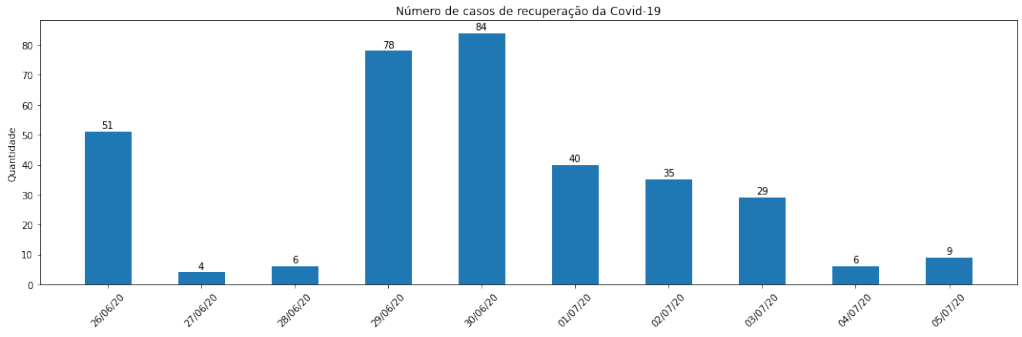

Figure 4. Casos recuperados.

Casos por faixa etária: No seguinte histograma é evidenciada a distribuição de porcentagem dos casos confirmados por faixa etária, é identificada a faixa mais afetada, os indivíduos de 41-50 anos de idade.

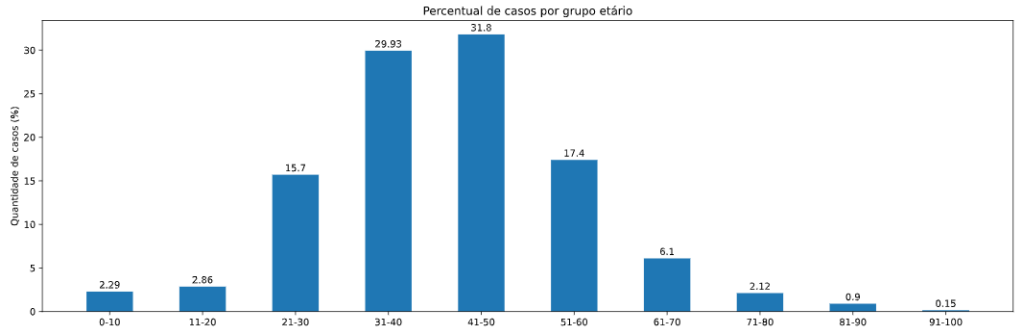

Figure 5. Casos por faixa etária.

Cumulativo de casos: O gráfico abaixo, desmonstra o cumulativo de casos notificados ao longo do tempo. No eixo x, estão localizadas as datas de notificação, que foi o atributo usado nesta tarefa e no eixo y, estão a quantidade de casos cumulativas. Para essa quantidade cumulativa ser calcula, foi usada um vetor que continha a quantidade de casos referente a cada dia de um determinado mês.

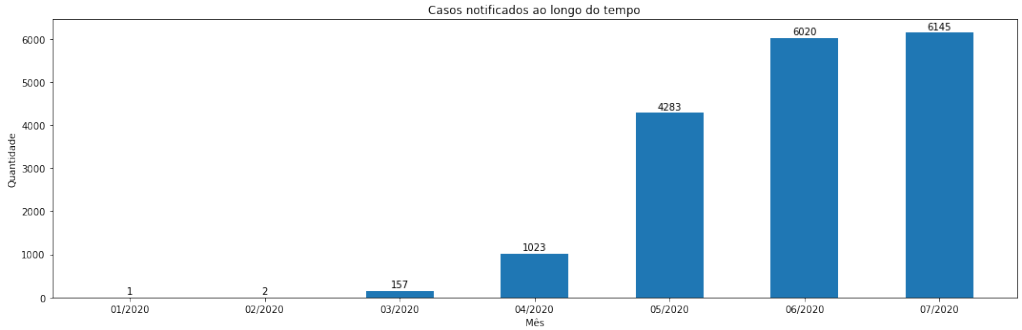

Figure 6. Cumulativo de casos.

Casos totais por idade: O gráfico abaixo, do tipo scatterplot, denota a idade(eixo x) versus o número total de casos naquela idade(eixo y). É perceptível a baixa correlação

na análise de Pearson entre ambos devido sua baixa intensidade. Apesar disso, é visível a tendência de aumento de casos de acordo com a idade, até o pico enquadrado na faixa etária próxima dos 35-55 anos, onde as variáveis se tornam inversamente proporcionais.

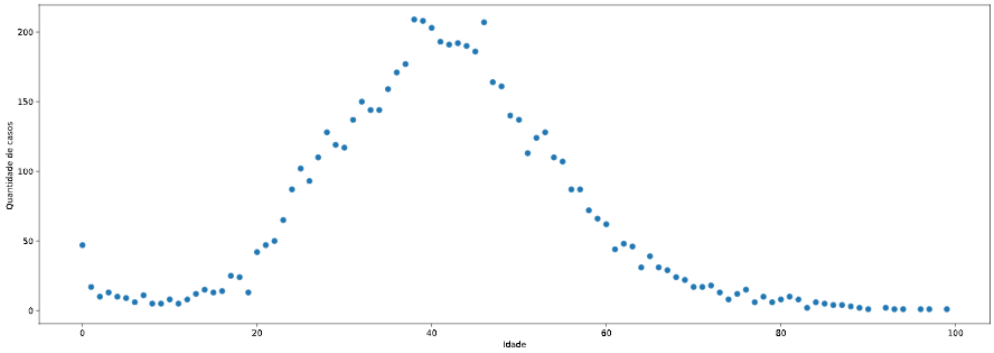

Figure 7. Quantidade de casos totais por idade

4. Tipos de Tarefas

A partir desses dados é possível fazer inúmeras análises, dentre elas uma classificação, mediante Aprendizado Supervisionado, de área de risco, tendo como alvo encontrar o Bairro com maior incidência, levando em consideração o atributo de conclusão equivalente a óbito, correspondendo ao falecimento de pacientes.

A partir dos campos válidos para a análise realizada, também é possível ver um fator que poder se levado em consideraçao para os pacientes de grupo de risco. A partir do atributo de conclusão levando em conta os pacientes que faleceram portando a doença, é possível verificar qual a faixa etária que se encontra na zona de risco, tendo como atributo alvo a idade para essa tarefa de regressão.

5. Referências

JACOFKSY, D. JACOFKSY, E. JACOFKSY M. Understanding Antibody Testing for COVID-19. 2020.

MatplotLib. User's Guide.Disponível em https://matplotlib.org/contents.html. Acesso em 05 de agosto de 2020.

NumPy. Installing Numpy. Disponível em https://numpy.org/install/. Acesso em 05 de agosto de 2020.

Pandas. Pandas Documentation. Disponível em https://pandas.pydata.org/docs/. Acesso em 05 de agosto de 2020.