Panorama da COVID-19 em Manaus: Análise Exploratória e Visualização de Dados

Dayvson Silva[1], **Fabrizio Honda**[1], **Hiago O. de Jesus**[1],
Jakson Protázio[1], **Jonatas Travessa**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Av. Darcy Vargas, 1.200 – Parque Dez de Novembro,
Manaus – AM, 69050-020

{ddss.snf19,fhf.lic17,hodj.lic,jpv.lic16,jtsdb.snf19}@uea.edu.br

Abstract. *Considering the pandemic state of COVID-19, the present study aims to realize an exploratory analysis and data visualization with the Manaus dataset, which contains data on the incidence of the disease in the city. This work consists of a technical report, required by the discipline of Artificial Neural Networks at the State University of Amazonas. The authors are computer academics and developed the work collaboratively, using the programming language Python 3.8 and Jupyter Notebooks. The GitHub repository link can be found in the Results.*

Resumo. *Considerando o estado de pandemia da COVID-19, o presente trabalho visa realizar uma análise exploratória e visualização dos dados com o dataset de Manaus, que contém dados de incidência da doença na cidade. Este trabalho consiste de um relatório técnico, requerido pela disciplina de Redes Neurais Artificiais da Universidade do Estado do Amazonas. Os autores são acadêmicos de Computação e desenvolveram o trabalho de forma colaborativa, com a linguagem de programação Python 3.8 e Jupyter Notebooks. O link do repositório em GitHub encontra-se nos Resultados.*

1. Introdução

Em 11 de março de 2020 foi declarado estado de pandemia pela Organização Mundial de Saúde (OMS) [Cucinotta e Vanelli 2020] devido à proliferação da doença do 'novo coronavírus', denominado COVID-19. Isto desencadeou a necessidade dos países de, temporariamente, fecharem suas fronteiras com outras nações e decretarem quarentena, vetando atividades escolares, comércio, shoppings, dentre outros, mantendo somente hospitais e postos de saúde. Em decorrência disso, afim de não comprometer completamente o calendário escolar anual, diversas escolas e universidades optaram pelo formato de ensino remoto, com aulas virtuais/online.

Na Universidade do Estado do Amazonas, o período de 2020_1 foi retomado em 2 de agosto[1], exclusivamente nesta nova modalidade de ensino. Em virtude disso, faz-se necessário a elaboração de novas estratégias de aprendizagem para o plano de ensino das disciplinas. Neste contexto, a docente da disciplina de Redes Neurais Artificiais adaptou o primeiro projeto avaliativo da disciplina a este novo modelo, requerindo que a turma se

[1]http://data.uea.edu.br/ssgp/noticia/1/65226-2.pdf

dividisse em equipes para realizar uma análise exploratória e visualização dos dados com o *dataset* de casos de COVID-19 em Manaus.

Este trabalho, portanto, descreve o processo de desenvolvimento deste projeto e seus resultados, em formato de relatório técnico, realizado por estudantes da disciplina. Na seção 2, a descrição do projeto é detalhada e as tecnologias utilizadas são informadas; na Seção 3, é abordado o desenvolvimento do projeto, com todas as atividades relacionadas (metodologia e resultados) e; na seção 4, as considerações finais.

2. Descrição do Projeto e Tecnologias

A requisição do projeto consiste em utilizar o *dataset* da COVID-19 em Manaus – disponibilizado pela prefeitura da cidade[2] – para resolver as atividades propostas de classificação, análise exploratória, visualização de dados e tipos de tarefas. Contendo 25 MB em formato *CSV (Comma-Separated Values)* e codificação ISO 8859-1, os dados devem ser manipulados por meio da linguagem de programação *Python* 3.6+ e bibliotecas *pandas* e *numpy*, em um ou mais *Jupyter Notebooks*. O repositório gerado deve estar disponível no *GitHub* e, por fim, a construção deste relatório técnico.

O projeto foi desenvolvido por cinco estudantes da disciplina – acadêmicos de Computação da Universidade do Estado do Amazonas –, cada um responsável por um tópico das atividades solicitadas (com exceção da visualização de dados, em que dois estudantes foram alocados). Para os gráficos, fez-se uso da biblioteca *matplotlib* e as atividades de programação nos *notebooks* foram realizadas tanto no *Google Colab* quanto na IDE *Visual Studio Code* com auxílio do gerenciador de pacotes *Anaconda*.

3. Desenvolvimento do projeto

O primeiro passo foi o *download* do *dataset*, que contém dados atualizados até 5 de agosto de 2020 – dia em que foi baixado. Em seguida fez-se uma análise, em que pôde-se observar que haviam 36.671 casos confirmados em Manaus, 17.191 em análise e 53.359 descartados. O registro mais antigo datava de 03/01/2020 e o mais recente em 05/08/2020, contudo, este primeiro não constava na base de dados como um caso confirmado. Oficialmente, o primeiro caso confirmado no município de Manaus foi noticiado pelos veículos de imprensa no dia 13/03/2020, porém, o primeiro caso registrado como confirmado data de 30/01/2020.

3.1. Visão Geral dos Casos Confirmados

Com o objetivo de considerar somente os casos confirmados, o processo de limpeza do conjunto de dados iniciou-se com a remoção dos atributos relativos às comorbidades, sintomas, etnia, profissão, outras datas que não a de notificação, origem e outros atributos não relacionados. Em seguida, excluiu-se todas as linhas nas quais haviam dados faltantes para os atributos remanescentes. Os registros que continham o atributo *classificação* com valores diferentes de *confirmado* foram removidos mantendo-se apenas os registros com o valor *confirmado* para o atributo *classificação*. Para fins de legibilidade, os nomes dos atributos foram renomeados e realizado o *casting* do atributo *idade* para um atributo do tipo inteiro. Por fim, executou-se a exportação do novo conjunto de dados como um arquivo no formato csv (*comma-separated values*), ocultando-se a indexação dos registros da base de dados mantendo como caracter delimitador a vírgula.

[2]https://covid19.manaus. am.gov.br/wp-content/uploads/Manaus.csv

3.2. Análise Exploratória

Após a limpeza dos dados, as atividades relacionadas a análise exploratória foram realizadas, descritas a seguir.

Quantos exemplos e atributos há na base de dados após a limpeza e organização? A etapa de limpeza resultou na remoção de 30 atributos, restando somente 6 atributos: idade, sexo, bairro, conclusão, data de notificação e tipo de teste. Quantos aos exemplos, restaram 6360 casos.

Qual a porcentagem de indivíduos recuperados em relação ao todo? A taxa de recuperados encontrada foi de 99,80%. No entanto, um resultado potencialmente afetado pela rigorosa limpeza da base de dados proposta pela atividade. É de conhecimento geral que a maioria das pessoas se recuperam da COVID-19, porém, a ausência de muitos dados na coluna 'conclusão' impediu a obtenção de um resultado mais preciso a respeito da taxa de recuperados.

Os casos acometeram mais indivíduos do sexo masculino ou feminino? A busca por infectados por gênero retornou os seguintes números: dos 6360 exemplos restantes após a limpeza, haviam 3605 casos entre mulheres e 2755 casos entre homens. Apesar do sexo feminino ter sido mais acometido, não há provas científicas suficientes para uma conclusão de que mulheres são mais susceptíveis a serem infectadas do que os homens.

Qual a média e desvio padrão de idade dos indivíduos que contraíram COVID-19? Qual o indivíduo mais jovem e o mais idoso a contraírem tal enfermidade? Quanto a infecção por idade, a maioria dos casos concentra-se em pessoas consideradas em uma idade economicamente ativa, ou seja, pessoas que de fato estão atuando no mercado de trabalho; a média de idade de 41 anos e o desvio padrão de 14,10 anos refletem essa realidade. Isso pode ser consequência do fato de que essas pessoas, ao saírem de casa para trabalhar, estão mais expostas do que aquelas que podem ficar em casa durante a quarentena.

Qual o bairro com maior incidência de casos? Quais os três bairros com maior incidência de casos recuperados? Sobre o número de casos por bairros, foi feita uma análise contando o número de casos por bairro e posteriormente ranqueando os bairros de acordo com o número de casos. Fazendo esta análise, a Cidade Nova foi o bairro com o maior número de registros, seguido por Flores e Tarumã. Há de se notar que foram confirmados casos em todas as regiões da cidade, e portanto pode-se sugerir que a transmissão por COVID-19 está ocorrendo em todas as zonas da cidade.

Quais os tipos de testes efetuados, segundo os dados? Indique os dados de maneira quantitativa e percentual. Quanto aos tipos de teste efetuados, a tabela 1 demonstra quantitativamente e percentualmente, os números obtidos.

Qual taxa de letalidade pode ser calculada a partir do conjunto de dados? Esta questão, em especial, foi realizada em dois momentos: antes e depois do processo de limpeza. Nesta primeira ocasião, chegou-se a uma taxa de letalidade de 5,55%. Porém, após a limpeza, o resultado indicou apenas 0,20%. Isso decorreu do fato de que, em muitos dos casos confirmados, a coluna _tipo_teste não indicava o tipo de teste utilizado para comprovar a condição de infectado pela COVID-19.

tipo_teste	realizados	porcentagem
RT-PCR	3693	58.07 %
TESTE RÁPIDO - ANTICORPO	1521	23.92 %
TESTE RÁPIDO - ANTÍGENO	1139	17.91 %
ECLIA IgG	4	0.06 %
ELISA IgM	3	0.05%

Tabela 1. Testes realizados

Qual o tipo de correlação, mediante coeficiente de correlação de Pearson, entre a idade e o número de casos? O coeficiente de correlação de Pearson indica o grau de correlação entre duas variáveis quantitativas exprimindo o grau de correlação através de valores situados no intervalo [-1,1]. Valores positivos indicam uma relação positiva entre as variáveis e quanto mais próximo de 1 mais linear é a relação, ou seja, quando o valor de uma variável aumenta os valores da outra variável também aumenta. Valores negativos indicam uma relação negativa entre as variáveis e quanto mais próximo de -1 mais inversa é a relação, ou seja, quando o valor de uma variável aumenta os valores da outra variável diminuem.

Foi solicitado tal cálculo entre a idade e o número de casos. Logo agrupou-se o número de casos por idade a fim de se obter uma informação a qual indicasse qual o tipo de correlação entre a idade e o número de casos. Mesmo antes de fazer o cálculo, é notável o fato de que a maioria dos casos ocorreram em pessoas economicamente ativas, e considerando a faixa etária dos registros da base de dados, a qual variava entre 0 a 99 anos de idade, esperava-se obter uma correlação fraca entre a idade e o número de casos, visto que a maioria dos casos não ocorria em crianças e idosos. Tal predição foi confirmada com o valor de -0,22% do coeficiente de correlação de Pearson, ou seja há uma correlação fraca e levemente negativa entre a idade e o número de casos.

3.3. Visualização de Dados

Conforme as solicitações exigidas no projeto, o próximo passo foi a realização de extrações dos dados para visualização e demonstração de relação, detalhados a seguir.

Histograma: quantidade de casos x bairros. Utilizando a base de dados que foi passada pelo processo de limpeza, obteve-se os nomes dos bairros. Em seguida, foram obtidos a quantidades de bairros que existiam no *dataset* e agrupou-se com a quantidade de casos notificados em cada bairro. Essa lista foi ordenada de acordo com a quantidade de casos notificados, assim apresentando dados dos bairros com mais casos. Após extrair os dez bairros com mais casos da COVID-19, o restante foi agrupado em uma lista chamada “Outros", podendo ser obervado no gráfico da figura 1.

Boxplot: idade dos casos confirmados. O gráfico abaixo demonstra a distribuição dos idades de casos confirmados em homens e mulheres. Pode-se observar que existem dados discrepantes como idades mais elevadas tanto para pessoas do sexo masculino quanto do sexo feminino. Também há outliers em idades de pessoas muito jovens para ambos os sexos (figura 2).

Gráfico de barras: últimos casos notificados. Para denotar os últimos casos foi feita uma ordenação crescente no dataset, com base nas datas de notificaçãoes dos casos.

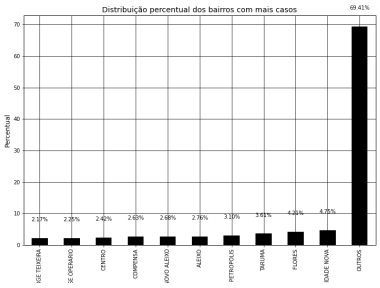

Figura 1. Histograma dos casos por bairros.

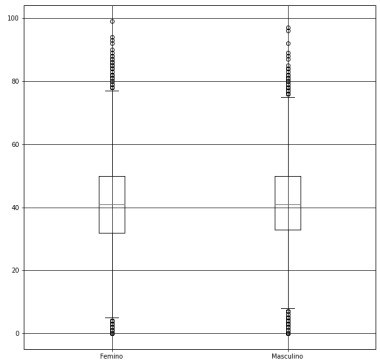

Figura 2. Boxplot dos casos confirmados.

Após, foi realizada um agrupamento das datas com base nas notificações dos casos e, em seguida, extraídos os últimos dez dias (figura 3).

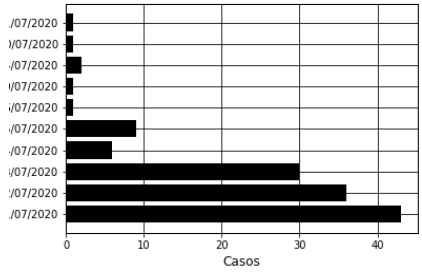

Figura 3. Gráfico de barras dos últimos casos notificados.

Gráfico de barras: casos recuperados Utilizando-se dos dados obtidos anteriormente, procurou-se identificar quais desses últimos casos conseguiram se recuperar, explicitado no gráfico da figura 4.

Histograma: casos por grupo etário Ao agrupar os dados por faixa etária (em décadas), obteve-se a classe de idades com as quais tiveram mais pessoas infectadas. Conforme aponta o gráfico abaixo, as faixas de idade entre 30 a 50 anos representam a maioria das pessoas que contraíram a COVID-19 (figura 5).

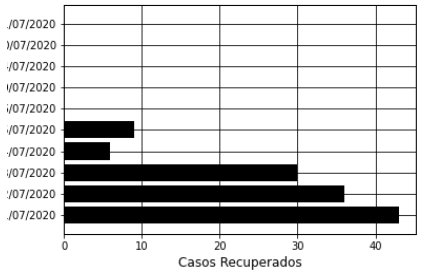

Figura 4. Gráfico de barras dos casos recuperados.

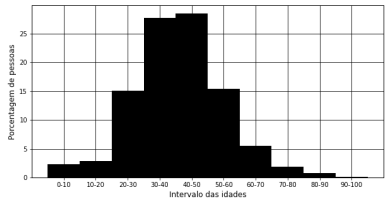

Figura 5. Histograma dos casos por grupo etário.

Gráfico de barras: casos notificados ao longo do tempo Como ao passar do tempo novos casos foram surgindo, foi possível montar um gráfico cumulativo demonstrando a quantidade total de casos notificados entre os meses de Março e Julho. Em Julho, mais de 6000 casos tinham sido notificados até então (figura 6).

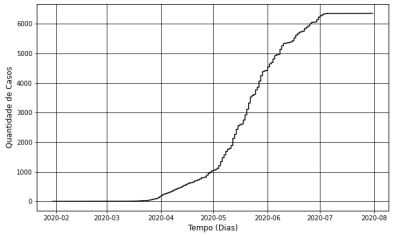

Figura 6. Graficos de barras dos casos notificados ao longo do tempo.

Scatterplot: idade x casos registrados. O número de casos registrados por idade pode ser agrupado e visualizado no gráfico de dispersão. Conforme cálculo realizado anteriormente, o coeficiente de correlação de Pearson para as variáveis idade e número de casos é de -0,22%. Logo, ao visualizar o gráfico e verificar o valor da correlação, pode-se visualizar que a correlação entre idade e número de casos é fraca. Sendo assim, não constatou-se uma tendência (figura 7).

3.4. Tipos de Tarefa

Propondo tarefa de classificação para a base de dados. Uma tarefa de classificação que pode ser realizada em cima dos dados limpos é: classificar um "paciente"como *Recuperado* da COVID-19 ou evoluiu para *Óbito*. Um dos problemas encontrados para realizar esta tarefa é o desbalenceamento de clases, pois enquanto a quantidade de "pacientes"recuperados é 6347, o número de pessoas com conclusão igual a Óbito chega a

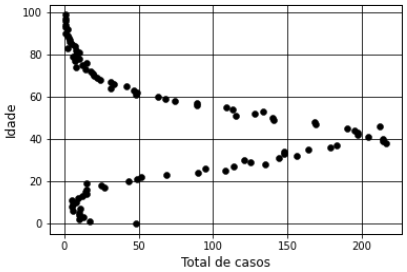

Figura 7. Scatterplot de idade por casos registrados.

13. Ao considerar o *dataset* não limpo, ou seja, com dados faltantes ou ainda somente as colunas _classificacao e _conclusao tem-se uma ideia mais real da taxa de mortalidade em Manaus (Total de registros confirmados: 36671, Quantidade de registros nulos: 23371, Toral de registros recuperados: 11264, Total de registros Óbito: 2036); isso decorre do fato da coluna _tipo_teste apresentar um número considerável de registros faltantes.

Uma possível forma de se contornar esse problema do desbalenceamento das clases seria selecionar outros atributos, pois observa-se que há bastante pessoas classificadas como Óbito, mas, por causa da limpeza e escolha desses atributos do *dataset* limpo, essa classe diminuiu consideravelmente. Outra sugestão para aumentar a quantidade de classes minoritária seria considerar mais registros com campos faltantes e depois adicionar a média do atributo para este campo ou encontrar outra medida para preencher tais campos.

Avaliação da tarefa de classificação. Como se trata de uma classificação binária, a acurácia é uma das principais métricas a ser considerada, porém, para este conjunto de dados talvez não reflita a total realidade por ele apresentar quantidades de classes desbalanceadas e ainda o auxílio de outras métricas torna possível a confirmação de que o modelo ou modelos propostos são eficientes. Então, a melhor maneira é adicionar outras métricas de válidação, como por exemplo a precisão que diz quantos acertos para uma classe (true positives) sobre o total de classificados como positivos (true positives + false positives). Aliados a estas métricas, acredita-se que as seguintes também tornariam a avaliação mais eficiente para o conjuto de dados considerado: *recall*, *f1-score* e também a *G-score* (bastante usada em modelos baseados em *datasets* com classes desbalanceadas).

Validação do/os modelo/os da tarefa de classificação. Para a validação do modelo, ou modelos caso se use mais de um para esta tarefa, pensou-se em criar dois experimentos: um usando a partição *holdout* e um outro usando o validação cruzada. O primeiro para identificar se com os dados é possível criar um bom classificador mesmo que tenha classes desbalanceadas e o segundo para garantir que o/os modelo/os possa/am ser treinado/os e testado/os com todos os registros e assim contornar o problema das classes desbalanceadas.

Propondo uma tarefa de regressão para a base de dados. Partindo do ponto de vista de gestores da área da saúde, uma das tarefas de classificação que podem ser feitas em cima dos dados limpos é quantidade de pessoas que serão infectadas por bairro, para assim concentrar mais recursos para esses pontos da cidade. O que garantiria que a quantidade de infectados tivessem cuidados propocionalmente, ou ainda caso a previsão fosse para dias a frente, poderiam ser tomadas medidas como o isolamento social nos bair-

ros com muitos casos previstos. Para esta tarefa, o primeiro passo seria o agrupamento de contagem dos atributos para os bairros, depois usar alguma metodologia para descobrir os melhores atributos para a tarefa como os metodos supervisionados. Contudo, acredita-se que principalmente os atributos sexo, idade, bairro e conclusao, são atributos que podem ter um peso maior para a tarefa por descreverem melhor grupos de pessoas de uma localidade. Uma outra tarefa de regressão pensada para este problema foi a previsão da idade dos pacientes com base em atributos como bairro, sexo, situação de conclusão do paciente, data de notificação e tipo de teste. Nesta segunda tarefa acreditasse que os atributos tipo_teste e df_notificação não tenham tanha relevância, pois eles não são características dos pacientes, mas sim dos testes que eles fizeram.

Qual tarefa de Aprendizado Não-Supervisionado poderia ser concebida neste contexto? Para esta questão, considerou-se o *dataset* pré-processado e a quantidade de atributos disponíveis. Uma tarefa de apredizado não-supervisionado que poderia ser feita é o agrupamento de perfis de pacientes mais parecidos, ou seja, agrupar os pacientes com base nas suas caracteríscas. Para realizar esse agrupamento, uma alternativa seria o uso do algoritmo KNN, assim possibilitando descobrir os perfis de quem se recuperou ou veio a óbito, por exemplo.

Quando se leva consideração não apenas o conjuto de dados limpos, mas também outras colunas deletadas no pré-processamento e supondo que elas tivessem menos dados faltantes e que não existisse a coluna _classificação uma outra tarefa não supervisionada seria a predição da classifiação do ẗacienteëm confirmado ou não com a COVID-19. Para esta tarefa seria usada a clusterização preditiva, assim como no trabalho de [Braga et al. 2019], usada para predizer os niveis de saúde em colônias de abelhas.

4. Considerações Finais

O trabalho em questão possibilitou que os estudantes trabalhem de forma colaborativa e explorassem mais o campo de Ciência de Dados, além de ser uma contribuição para análise dos casos de COVID-19. O código-fonte do repositório em GitHub pode ser encontrado no link: https://github.com/userddssilva/Analise-COVID19-PP1-RNA2020.1

Referências

[Braga et al. 2019] Braga, A. R., Silva, D. A., Nobre, J. S., Freitas, B. M., and Gomes, D. G. (2019). Definindo e predizendo níveis de saúde de colônias de abelhas via clusterização e classificação. In *Anais Principais do XXXIV Simpósio Brasileiro de Banco de Dados*, pages 241–246. SBC.

[Cucinotta and Vanelli 2020] Cucinotta, D. and Vanelli, M. (2020). Who declares covid-19 a pandemic. *Acta bio-medica: Atenei Parmensis*, 91(1):157–160.