Redes Neurais Artificiais:
Projeto Prático 1

Ian Gustavo Alves Pessoa[1], **Levi da Silva Lima**[2], **William Azevedo da Silva**[3], **Daniel Akio Chen**[4], **Francisco Elio Parente Arcos Filho**[5]

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

{igaps.eng17, ldsl.eng16, wads.eng16, dac.eng17, fepaf.eng}@uea.edu.br

Abstract. *This work aims to make a study on the COVID-19 case database in the minicipio de manaus taking into account a period from 01/04/2020 to 07/31/2020. For this, the development tools present for Python were used with the help of Google Colab development environment, which gathers devices for analyzing large amounts of data. The auxiliary tools used were pyplot, plotly, pandas and spicy which were used for specific calculations, data handling and graph plotting.*

Resumo. *Este trabalho tem por objetivo fazer um estudo sobre a base de dados de casos de COVID-19 no município de Manaus levando em consideração um período que vai de 01/04/2020 à 31/07/2020. Para tal foram utilizadas as ferramentas de desenvolvimento presentes para Python com o auxilio do ambiente de desenvolvimento Google Colab que reúne aparatos para analise de grande quantidade de dados. As ferramentas auxiliares utilizadas foram pyplot, plotly, pandas e spicy para cálculos específicos, manipulação de dados e plote de gráficos.*

1. Introdução

Este trabalho tem como objetivo analisar a base de dados de casos de covid-19 no municipio de Manaus. A partir dos casos listados no dataset será feito um pre-processamento (limpeza) sobre os dados ruidosos para posterior analise. O trabalho será divido em 3 etapas principais, quais sejam: visao geral dos casos confirmados com análise exploratória, visualização dos dados e por fim a susgestao de tarefas de aprendizado para o problema. Por fim os topicos de tarefas seram todos feitos usando a ferramente de desenvolvimento google colab com auxilio do python e algumas ferramentas adicionais para gerar graficos.

2. Fundamentação Teórica

Python é uma linguagem de programação de alto nível,interpretada, de *script*, imperativa, orientada a objetos, funcional, de tipagem dinâmica e forte. Foi lançada por Guido van Rossum em 1991. Atualmente possui um modelo de desenvolvimento comunitário, aberto e gerenciado pela organização sem fins lucrativos Python Software Foundation. Apesar de várias partes da linguagem possuírem padrões e especificações formais, a linguagem como um todo não é formalmente especificada.

Dentre as diversas bibliotecas presentes no *python*, para esse projeto vale ressaltar: *pandas*, *matplotlib*, *plotly*, *math*, *scipy*, *datetime* que são usadas para fazer cálculos específicos e também para construir imagens e referencias interativas em forma de gráficos sobre um conjunto de informações.

Por fim o ambiente de codificação Google Colab que se reuniu como um coletivo em 1977, usando pela primeira vez o nome Green Corporation, e inicialmente recebeu uma bolsa *nea workshop* através do Center for New Art Activities, Inc. Um pequeno sem fins lucrativos formado em 1974 que reúne todas essas ferramentas de forma a permitir o uso sem qualquer restrição por parte dos usuários.

Git é um sistema de controle de versões distribuído, usado principalmente no desenvolvimento de software, mas pode ser usado para registrar o histórico de edições de qualquer tipo de arquivo. GitHub é uma plataforma de hospedagem de código-fonte com controle de versão usando o Git.

3. Desenvolvimento

3.1. Visão de Casos Confirmados

No que tange as definições iniciais relacionadas ao *dataset* de casos de Covid-19 em Manaus é importante ressaltar algumas considerações iniciais sobre os dados, levando em consideração que a data de coleta da base de dados para estudo foi no dia 31/07/2020. Sendo assim destaca-se:

- Cada exemplo da base de dados é descrito por 36 valores distintos, sendo eles: '_idade', '_faixa etária', '_sexo', '_bairro', '_classificacao', '_comorb_renal', '_comorb_diabetes', '_comorb_imuno', '_comorb_cardio', '_conclusao', '_dt_notificacao', '_taxa', '_dt_evolucao', '_raca', '_dt_sintomas', '_criterio', '_tipo_teste', '_sintoma_garganta', '_sintoma_dispneia', '_sintoma_febre', '_sintoma_tosse', '_sintoma_outros', '_etnia', '_profiss_saude', '_srag', '_se_notificacao', '_distrito', '_bairro_mapa', '_comorb_respiratoria', '_comorb_cromossomica', '_comorb_hepatica', '_comorb_neurologica', '_comorb_hemato', '_comorb_obessidade', '_origem', '_evolução'
- Segundo a base de dados, em Manaus há cumulativamente 36671 casos confirmados
- O período de tempo que a base de dados abrange vai de 01/04/2020 a 31/07/2020

Para uma visualização mais concreta e precisa, foi feita uma limpeza na base de dados de casos de COVID-19, entre elas foi feita a remoção das colunas de : _comorb_renal', '_comorb_diabetes', '_comorb_imuno', '_comorb_cardio', '_comorb_respiratoria', '_comorb_cromossomica', '_comorb_hepatica', '_comorb_neurologica', '_comorb_hemato', '_comorb_obessidade', '_sintoma_garganta', '_sintoma_dispneia', '_sintoma_febre', '_sintoma_tosse', '_sintoma_outros', '_etnia', '_raca', '_profiss_saude', '_dt_evolucao', '_dt_sintomas', '_origem', '_evolução', '_criterio'.

3.1.1. Análise exploratória

Com a base de dados limpa, pode-se responder alguns questionamentos sobre a estrutura e a disposição dos dados. Apos a limpeza e organização da base de dados é possível se

afirmar que do total restaram 36671 exemplos sendo estes descritos por 13 atributos cada, implicando assim em uma queda aproximada de 65% com relação a quantidade original de dados o que leva a concluir que havia uma quantia considerável de dados ruidosos. Sendo assim, a partir desse ponto sempre que for mencionado base de dados essa já corresponde aos dados consistentes filtrados.

Para se ter uma sensibilidade percentual dentre os casos de COVID-19 dados como recuperados e não recuperados, foi feita uma análise sobre a base de dados obtendo-se o resultado de que 30.72% de todos os casos presentes foram dados como recuperados. Vale ressaltar que a maior quantidade de casos acometidos de COVID-19 foram em mulheres. A Figura 1 demonstra graficamente o resultado.

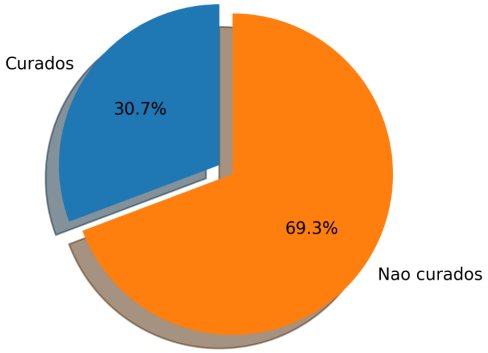

Figura 1. Gráfico de relação percentual entre recuperados e não recuperados

Facilmente percebe-se que não basta um simples calculo percentual para se ter um conhecimento solido sobre os dados, então para um melhor conhecimento sobre a disposição das informações foi feito um estudo estatístico, sobre a idade dos indivíduos que contraíram COVID-19, que envolvem a media e o desvio padrão pois eles permitem ter, respectivamente, uma visão geral da base de dados e também ter uma noção do grau de variação entre o conjunto de dados. Sendo assim, obteve-se o valor de 43.0 para media que indica que no geral as pessoas que contraíram a doença tem 43 anos e 16.92 para o desvio padrão que mostra que os valores podem se diferenciar da média por 17 anos indicando um possível intervalo de 26 a 60 anos para os indivíduos que contraíram a doença. Para concluir essa etapa da análise vale ressaltar qual é o verdadeiro intervalo de idade entre os indivíduos listados na base de dados, sendo assim é possível indicar que o individuo mais velho a contrair a doença tinha 120 anos e o mais novo 0 anos (criança recém nascida).

Partindo agora para um estudo que leva mais em consideração a disposição espacial das pessoas que moram em Manaus vale citar que o bairro com maior incidência de casos confirmados foi o bairro CIDADE NOVA com um total de 2008 casos e em contra partida a isso, os 3 bairros com maior quantidade de pessoas dadas como RECUPERADAS foram: CIDADE NOVA, FLORES e CENTRO. Dentre os casos confirmados há ainda aqueles que não fizeram nenhum tipo de exame, no entanto, dentre os que fizeram constata-se que estes foram de 5 tipos diferentes, sendo eles: ECLIA IgG, ELISA IgM, RT-PCR, TESTE RÁPIDO - ANTICORPO, TESTE RÁPIDO - ANTÍGENO. Para se ter uma maior sensibilização sobre os testes feitos tanto de forma quantitativa como percentual observe os gráficos na Figuras 2 e na Figura 3 que mostram a disposição dessas

informações.

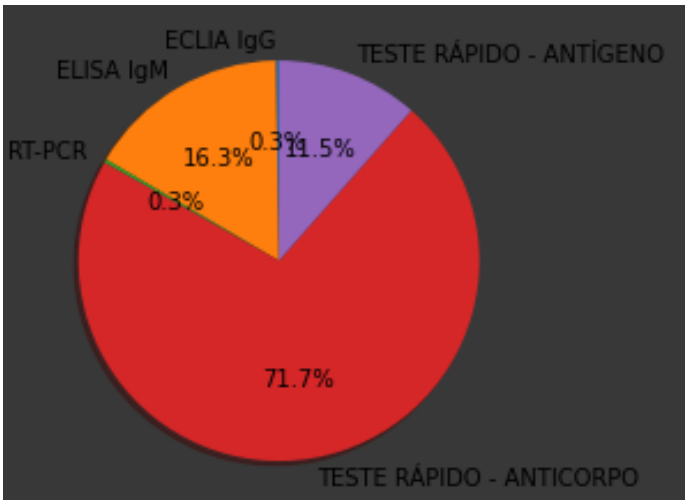

Figura 2. Gráfico de percentual de cada teste que foi feito

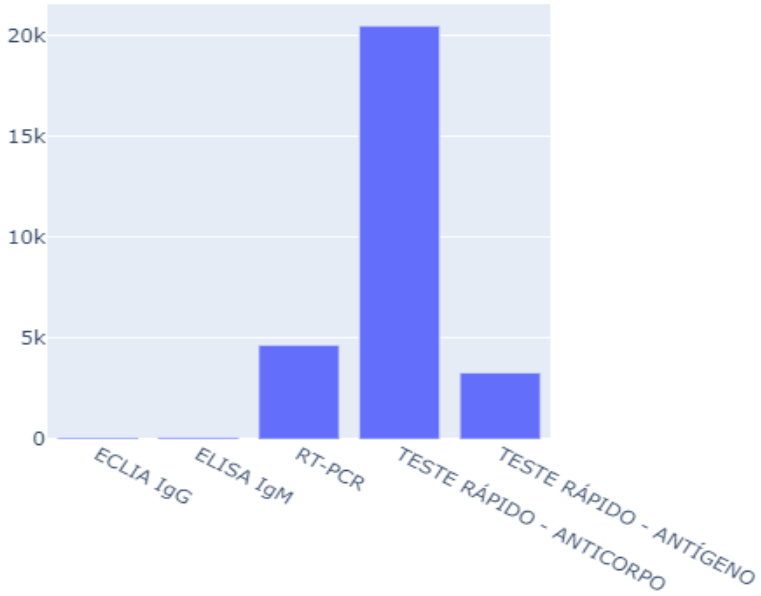

Figura 3. Gráfico de quantidade de cada teste que foi feito

Para finalizar essa etapa de análise exploratória dos dados é importante apontar que a taxa de letalidade foi de 5.5%, que pode ser confirmado, o que indica um baixo valor percentual. Por fim o coeficiente de correlação de Pearson que visa medir as relações entre variáveis e o que elas representam foi de aproximadamente -0.3309 o que indica um correlação fraca e negativa ou inversa, indicando que as variáveis possuem baixa correlação. Como mostrado na Figura 4, os casos possuem uma distribuição normal.

3.2. Visualização de Dados

Para se ter uma noção da disposição de casos de COVID-19 em Manaus, foi feito um estudo sobre a distribuição percentual de ocorrências, levando em consideração um agrupamento por bairros. Tendo isso em vista chegou-se ao seguinte gráfico da Figura 5 que

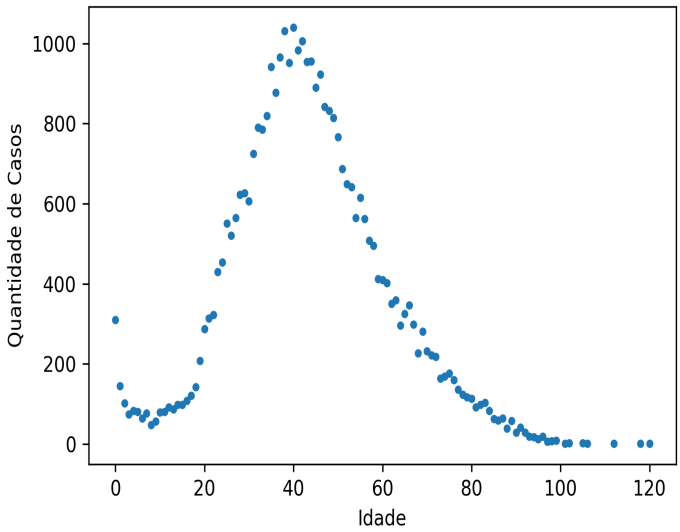

Figura 4. Gráfico de Distribuição por idade

demonstra os 10 bairros com maior percentual de afetados (eixo x) e seus respectivos percentuais (eixo y), todos os demais bairros encontram-se juntos agrupados pelo rótulo de 'OUTROS'. Observa-se que existe certa uniformidade entre os bairros com maior índice de contaminados.

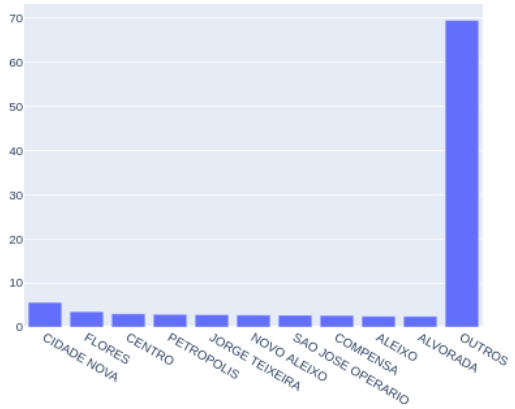

Figura 5. Gráfico de percentual nos 10 bairros de maior ocorrência

Em adição a isso percebe-se que a variação entre os dados é de certa forma bem significante, para exemplificar isso foi feito o boxplot de casos confirmados por idade que estão mostrados na Figura 6 e é perceptível a presença de *outliers* (valores discrepantes) de forma expressiva na faixa 82 a 100 anos. Em conjunto com a Figura 4, nota-se que as caudas da distribuição são largas e a dispersão é baixa visto pelo tamanho da caixa representando os quartis. Os *outliers* mostram que mesmo os mais idosos, acima de 100 anos e que são minoria, e as crianças mais novas, não foram privados do contágio, visto que de acordo com a Figura 4, existe um pico para os recém nascidos.

Com um intuito de sensibilização do nível de contaminação do vírus, foi feito

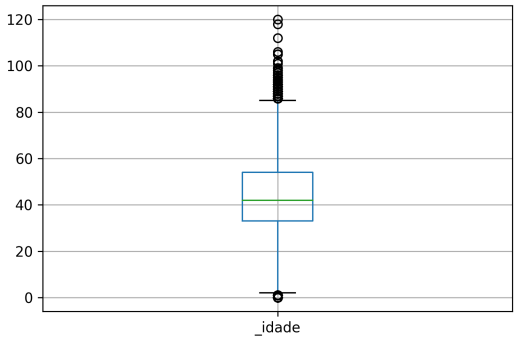

Figura 6. Boxplot da idade dosS casos confirmados

um gráfico, figura 7, que mostra a quantidade de novos casos confirmados nos últimos 10 dias contidos na base de dados, e também, para critério comparativo, foi feito um segundo gráfico similar ao da figura 7 mas, levando em consideração a quantidade de pessoas recuperadas. O mesmo é representado pela figura 8.

Como é possível constatar, a ocorrência de novos casos é muito mais recorrente que a de casos recuperados. Isso demonstra que a incidência do vírus tende a aumentar cada vez mais. É possível perceber também que todos os dias ouve novas notificações de casos enquanto de recuperados não.

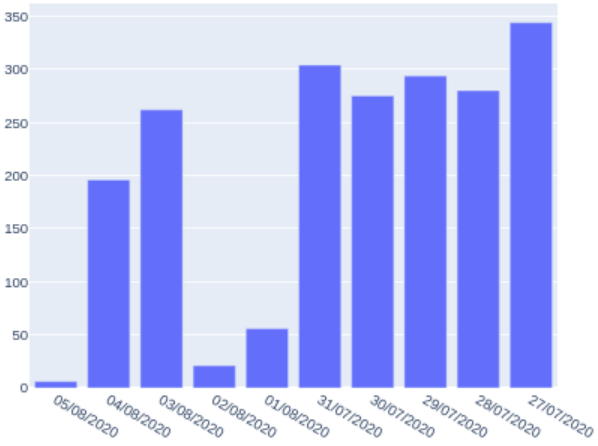

Figura 7. Número de novos casos por dia, considerando os 10 últimos dias existentes na base de dados

Visando analisar a quantidade de casos do vírus em relação as idades, foi feito um histograma, apresentado na figura 9, no qual é apresentado a relação Porcentagem x Faixa etária. Pelo histograma (figura 9), notamos que a quantidade de casos é baixa para as faixas etárias mais jovem e vai crescendo até 31 a 40 e 41 a 51. Depois disso, começa a cair, gerando assim, uma figura muito parecida com uma distribuição normal. Isso mostra que pessoas com idade de trabalho, provavelmente as que tiveram que sair de suas casas, ficaram mais vulneráveis ao vírus.

Para efeito demonstrativo o gráfico da figura 10 mostra a curva da quantidade total

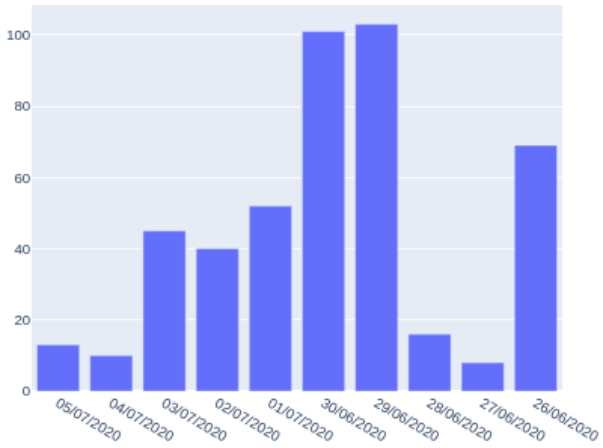

Figura 8. Número de novos casos recuperados, considerando os 10 últimos dias existentes na base de dados

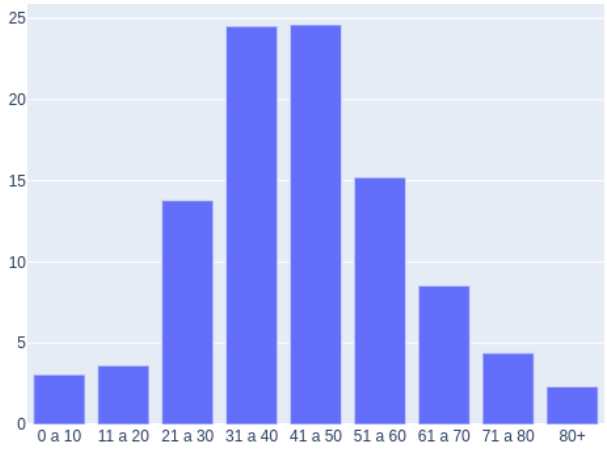

Figura 9. Histograma Porcentagem de Casos x Faixa etária

de casos ao longo do tempo. Ao observar esse gráfico facilmente nota-se que a curva é ascendente com uma inclinação bem acentuada mas que por volta do mês 06 sofreu um pequeno desvio e seus valores começaram a diminuir (ainda sim é muito inclinada) semelhante a uma função logarítmica mas que ainda não chegou em seu ponto de equilíbrio e que mostra um possível aumento ainda nos meses seguintes.

Já no gráfico 11, podemos ver um gráfico que denota a idade relacionada ao total de casos registrados para aquela certa idade, ou seja, usando os dados brutos. Comparando ao gráfico 4, os dois são muito semelhantes, em ambos os gráficos, existe a tendência que os mais afetados são aqueles com aproximadamente 40 anos, mas também houve um pequeno pico nos recém-nascidos. Ou seja, o trabalhador médio que é em sua maioria na faixa de 40 anos estava mais sujeito a contaminação, nos extremos, a intensidade é menor porém ela é uniformemente distribuída e os recém nascidos foram bastante vulneráveis.

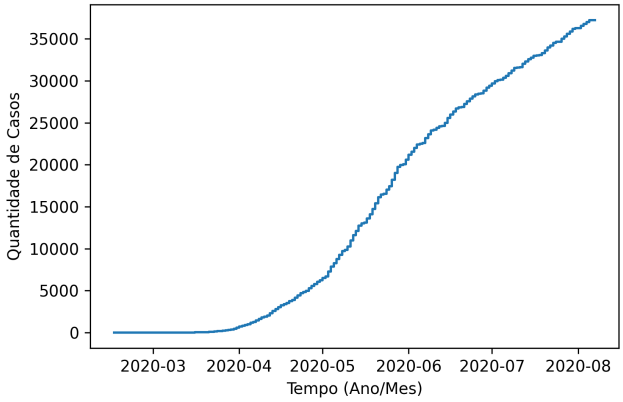

Figura 10. Gráfico cumulativo de casos notificados ao longo do tempo

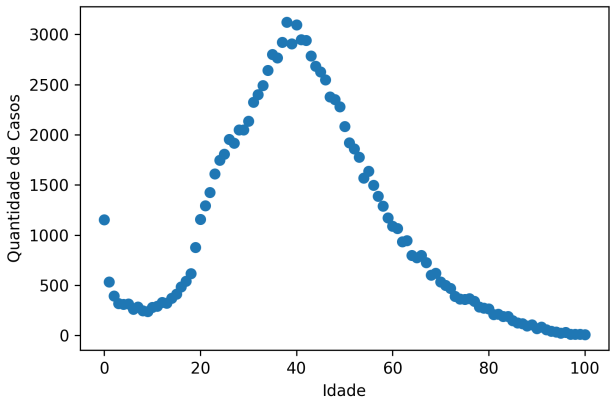

Figura 11. Gráfico da idade versus o número total de casos registrados. Todos os casos registrados

3.3. Tipos de Tarefas

3.3.1. Tarefa de Classificação mediante Aprendizado Supervisionado

Uma tarefa de classificação que poderia ser usada é levar em consideração informações como: idade, comorbidades (diabetes, hipertensão), classificação (confirmado ou não) e a partir disso poder prever uma possível situação final para o individuo (recuperado ou óbito), para tanto pode-se utilizar algoritmos de aprendizado supervisionado como rede neural. Nesse contexto o atributo alvo para a rede neural seria 'conclusão' mostrando assim se o individuo se recuperou ou veio a óbito. Mas não basta apenas criar um modelo, treina-lo e por em pratica, é necessário avaliar a qualidade do modelo produzido e para isso há algumas métricas capazes de aferir tais características. Uma métrica que caberia a esse problema é o *F1 Score* que é uma média harmônica entre precisão, ela é muito boa quando você possui um *dataset* com classes desproporcionais (esse caso em particular é um desses), e o modelo não emite probabilidades (também se encaixa aqui). Obs: Isso não significa que não possa ser usada com modelos que emitem probabilidades, tudo depende do objetivo da tarefa de *machine learning*.

Precision também poderia ser usado nessa caso já que ele basicamente analisa o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe (positivos verdadeiros) e divide pela soma entre este número, e o número de exemplos classificados nesta classe, mas que pertencem a outras (falsos positivos). O *recall* que é bem similar ao que o *precision* faz.

Para a validação dos dados pode-se utilizar métodos como *Holdout*, *Cross-Validation* e *Leave-one-out* que permitem fazer a verificação do modelo de forma mais empírica.

3.3.2. Tarefa de Regressão mediante Aprendizado Supervisionado

Uma possível tarefa de regressão utilizando aprendizado supervisionado seria calcular a porcentagem de chance de uma pessoa recuperar-se, semelhante a questão anterior, porém, retornando a porcentagem de chance considerando um novo paciente. Os atributos preditores seriam os mesmo do método de classificação citado no exemplo a cima: idade, comorbidades, classificação (quanto a confirmação do vírus).

Para a avaliação de desempenho, poderia ser utilizado a raiz quadrada do erro médio, que basicamente verifica o quão perto estão os dados da função gerada pela regressão.

3.3.3. Tarefa de Aprendizado Não-Supervisionado

No contexto de aprendizagem não supervisionada, poderia ser aplicado à este *dataset*, uma clusterização dos dados, visando observar possíveis grupos de padrões (clusters) mostrando novas formas de observar os dados. Um possível algoritmo de clusterização a ser utilizado é o *K-Means*, que agrupa os dados em *K Clusters*.

4. Código Fonte

Todo o trabalho foi versionado utilizando a ferramenta Git e está hospedado na plataforma Github no repositório endereçado em `https://github.com/levidasilvalima/rna-pp1`

5. Conclusão

Ao realizar este trabalho, foi possível aplicar vários conceitos sobre tratamento de dados, análise de dados (estatística) e exibição de dados (gráficos), aprendidos em sala de aula e durante estudos pessoais, o que certamente contribuiu para a formação profissional dos membros da equipe.

Vale ressaltar que trabalhar com um problema atual e importante (COVID-19) foi estimulante para os membros da equipe, que ficaram gratificados em ver os resultados e comparar com os resultados divulgados pelos meios de comunicação.

Referências

[Faceli et al. 2011] Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.