

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Inteligência Artificial e Aprendizado de Máquina

Cleiton Neves Santos

**PRIORIZAÇÃO DE INDIVÍDUOS PARA TESTAGEM DE COVID-19 EM SITUAÇÃO
DE ESCASSEZ DE TESTES**

Belo Horizonte
Agosto

Cleiton Neves Santos

**PRIORIZAÇÃO DE INDIVÍDUOS EM ESCASSEZ DE TESTES COM AUXÍLIO DE
APRENDIZADO DE MÁQUINA**

Trabalho de Conclusão de Curso apresentado
ao Curso de Especialização em Inteligência
Artificial e Aprendizado de Máquina, como
requisito parcial à obtenção do título de
Especialista.

Belo Horizonte

Agosto

SUMÁRIO

1. Introdução.....	4
2. Descrição do Problema e da Solução Proposta.....	4
3. Canvas Analítico.....	5
4. Coleta de Dados.....	6
5. Processamento/Tratamento de Dados.....	7
6. Análise e Exploração dos Dados.....	8
7. Preparação dos Dados para os Modelos de Aprendizado de Máquina.....	10
8. Referências.....	11

1. Introdução

Nos últimos anos o mundo vem enfrentando uma pandemia da doença COVID-19, causada pelo vírus SARS-CoV-2. Com isso, a atenção de diversas áreas do conhecimento se voltaram para a procura de meios de conter a propagação da doença e os danos que ela causa. Uma das áreas onde a computação pode contribuir é na assistência à decisão médica.[1] Tendo isso em vista, este relatório tem o objetivo de propor uma estratégia de decisão para priorizar, e através disso agilizar, a testagem de pacientes com suspeita de Covid-19 em um cenário de escassez de testes disponíveis.

Existe um artigo [2] propondo uma estratégia de priorização de testagem em pandemia de Covid-19 por meio do uso de aprendizado de máquina. Porém, este relatório se diferencia do artigo encontrado à medida que os fatores para priorização na pesquisa citada são demográficos e sintomáticos (quais sintomas e características demográficas devem ser priorizadas para testagem), e seus resultados poderiam ser utilizados até mesmo concomitantemente com as ideias deste relatório.

Este relatório se diferencia dos trabalhos disponíveis na plataforma Kaggle a medida em que o modelo construído entregará não uma classificação (1 ou 0) mas sim uma probabilidade do paciente estar na classe “Positivo para Covid-19”, probabilidade essa que servirá de auxílio à decisão. O relatório sugere a estratégia de priorizar a testagem de indivíduos classificados com baixa confiança (valores próximos de 50% de chance de pertencimento à classe “Positivo para Covid-19”).

2. Descrição do Problema e da Solução Proposta

A pandemia de Covid-19 motivou diversos estudos sobre o impacto da testagem e rastreio de contatos como estratégia de contenção do avanço da doença em questão. As pesquisas indicam que a redução de tempo entre o aparecimento de sintomas e o diagnóstico pode reduzir significativamente o número de casos e mortes [3][4]. Porém, em muitas situações, os recursos materiais de cada país impedem a implementação da estratégia ideal [5][6]. Apesar de resultados positivos em artigos sobre o diagnóstico de Covid-19 usando modelos de predição [7], o que








indicaria a possibilidade de substituição total dos testes convencionais, preferiu-se neste relatório empregar o modelo apenas como auxílio em caso de escassez de testes. Apresentando assim, uma outra alternativa de emprego da tecnologia.

Em face da conclusão de que a demora da confirmação de casos tem um considerável impacto na propagação da doença, este relatório tem o objetivo de apresentar uma estratégia alternativa que agilize a testagem em situação de escassez de testes. A proposta é utilizar dados sobre sintomas da doença para a criação de um modelo de aprendizado de máquina, que irá classificar o indivíduo como positivo ou negativo para Covid-19 com uma confiança associada. A proposta é destinar os testes para os casos onde a confiança do modelo se mostrou menor, reduzindo as chances de erro.

3. Canvas Analítico

Software Analytics Canvas

Project: _____

<p> 1. Question</p> <p><i>What is it that we want to know about the software / processes / usage / organization / etc.?</i></p> <p>Quem priorizar em momentos de escassez de testes para certa doença transmissível? Como dados podem auxiliar nessa priorização.</p>	<p> 2. Data Sources</p> <p><i>Which data can possibly answer our question? What information do we need?</i></p> <p>Para tentar responder o questionamento, será necessário obter dados com informações sintomáticas e de positivo/negativo de uma doença transmissível.</p>	<p> 3. Heuristics</p> <p><i>Which assumptions do we want to make to simplify the answer to our question?</i></p> <p>Consideraremos que os resultados nas bases de dados não são "falsos positivos" ou "falsos negativos". O que é razoável considerando que a análise está sendo feita a posteriori e resultados errados seriam ajustados nas bases de dados.</p>	<p> 4. Validation</p> <p><i>What results do we expect from our analysis, how are they reviewed and presented in an understandable way?</i></p> <p>O resultado esperado é a criação de um modelo que auxilie na priorização de testagem de suspeitos de alguma doença. O objetivo é que essa seja uma ferramenta útil em situação de carência de testes.</p>
<p> 5. Implementation</p> <p><i>How can we implement the analysis step by step and in a comprehensible way?</i></p> <p>A Implementação será feita da seguinte forma:</p> <ul style="list-style-type: none"> - Coleta, limpeza e tratamento de dados pertinentes; - Análise de que algoritmo será o mais indicado para a criação de um modelo de classificação, para o caso específico; - Após a criação e refinamento do modelo, aplicar para uma base de dados e ordenar as classificações de acordo com a probabilidade de classificação correta; - Por fim, permitir a visualização dos resultados com o intuito de auxiliar na priorização da testagem de indivíduos para Covid-19. 		<p> 6. Results</p> <p><i>What are the main insights from our analysis?</i></p>	<p> 7. Next Steps</p> <p><i>What follow-up actions can we derive from the findings? Who or what do we need to address next?</i></p>

Software Analytics Canvas v1.0 designed by Markus Harrer. Visit <https://www.feststelltaste.de/software-analytics-canvas/> for more information. CC BY-SA 4.0

(Os itens 6 e 7 serão terminados nas seguintes partes do projeto).

4. Coleta de Dados

Para a criação do modelo de classificação, é importante a alimentação da base com dados que, em uma primeira análise, sejam relevantes para o tema. Pois isso será um aspecto relevante na qualidade do modelo de predição a ser criado.

Será utilizado um conjunto de dados sobre sintomas de pacientes com suspeita de covid-19 disponibilizado na internet pela Organização Mundial de Saúde (OMS). Os dados foram obtidos através da comunidade para ciência de dados e repositório aberto de dados, Kaggle, e são referentes a 2020.

Mais informações sobre o conjunto de dados:

Nome do dataset: Symptoms and Covid presence. Descrição: As primeiras colunas representam diferentes sintomas da Covid-19, a última indica se a doença foi identificada. Contém 5434 registros na base. Link: Kaggle Dataset (acesso em 20/04/2022)		
Nome do Atributo	Descrição	Tipo
Breathing Problem	Se o indivíduo apresenta problema de respiração.	Categórico; binário.
Fever	Se o indivíduo apresenta febre.	Categórico; binário.
Dry Cough	Se o indivíduo apresenta tosse seca.	Categórico; binário.
Sore throat	Se o indivíduo apresenta dor de garganta.	Categórico; binário.
Running Nose	Se o indivíduo apresenta nariz escorrendo.	Categórico; binário.
Asthma	Se o indivíduo apresenta sintomas de asma.	Categórico; binário.
Chronic Lung Disease	Se o indivíduo apresenta doença pulmonar crônica.	Categórico; binário.
Headache	Se o indivíduo apresenta dor de cabeça.	Categórico; binário.
Heart Disease	Se o indivíduo apresenta doença coronária.	Categórico; binário.
Diabetes	Se o indivíduo apresenta diabetes.	Categórico; binário.
Hyper Tension	Se o indivíduo apresenta hipertensão.	Categórico; binário.
Fatigue	Se o indivíduo apresenta cansaço.	Categórico; binário.
Gastrointestinal	Se o indivíduo apresenta	Categórico; binário.

	problemas gastrointestinais.	
Abroad travel	Se foi ao exterior nos últimos 14 dias.	Categórico; binário.
Contact with COVID Patient	Se teve contato com pessoas infectadas.	Categórico; binário.
Attended Large Gathering	Se o indivíduo participou de aglomerações.	Categórico; binário.
Visited Public Exposed Places	Se visitou locais públicos.	Categórico; binário.
Family working in Public Exposed Places	Se alguém do núcleo familiar do indivíduo trabalha em local público.	Categórico; binário.
Wearing Masks	Se o indivíduo faz uso de máscaras de proteção.	Categórico; binário.
Sanitization from Market	Se o indivíduo se sanitiza ao retornar do mercado.	Categórico; binário.
COVID-19	Se o indivíduo testou positivo para Covid-19	Categórico; binário.

5. Processamento/Tratamento de Dados

Para o processamento de dados e criação do modelo será utilizada a linguagem de programação *Python* em associação, principalmente, com a biblioteca de manuseio e análise de conjunto de dados chamada *pandas*.

Foi checado a presença de dados faltantes ou nulos na base, como não foi encontrado não foi necessária a aplicação de nenhuma técnica de preenchimento de valores faltantes ou remoção de dados. Todas as variáveis do conjunto de dados são categóricas e não numéricas. Sendo assim, optei por transformar as variáveis em numéricas pois são compatíveis com uma maior variedade de algoritmos de aprendizado (Figura 1).

Figura 1 - Categorias numéricas

```

#Função de tratamento de dataset

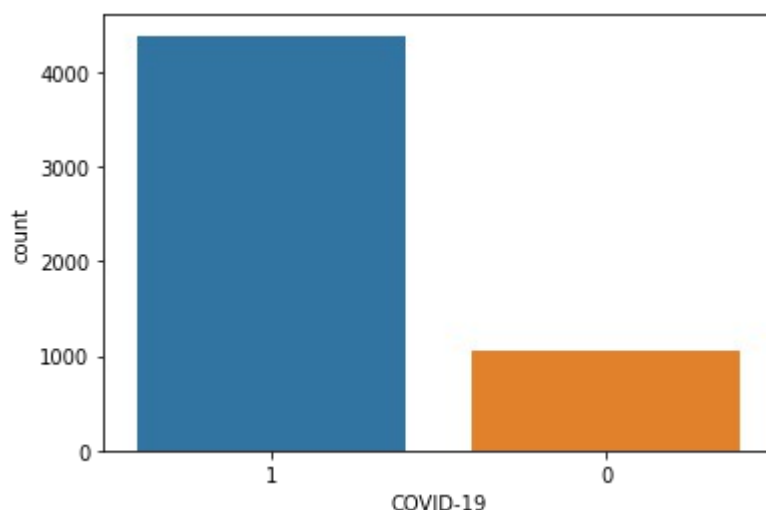
def tratamento(df):
    df.replace('Yes', 1, inplace=True);
    df.replace('No', 0, inplace=True);
    return df;

```

[2] ✓ 0.2s

Foi observado que o conjunto de dados é desbalanceado, com a categoria “Yes” para “COVID-19” tendo 3332 registros a mais (Figura 2). Para minimizar o viés do modelo a ser criado, será necessária a aplicação de alguma técnica de balanceamento nas próximas etapas do processo.

Figura 2 – Base Desbalanceada



E por fim, foram removidas as variáveis correspondentes às colunas “Wearing Masks” e “Sanitization from Market”. Pois essas colunas apresentavam um mesmo valor em todas as entradas, sendo assim dispensáveis para a criação do modelo de classificação.

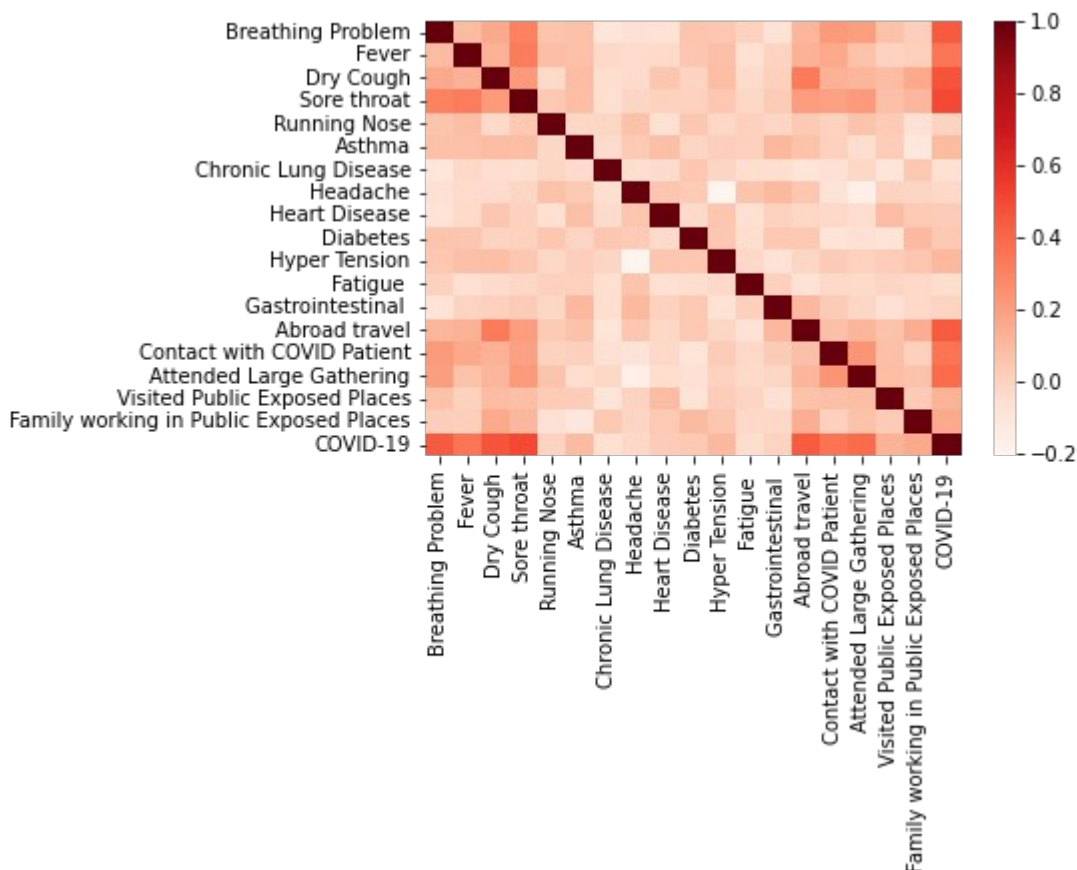
6. Análise e Exploração dos Dados

Ao fazer uma análise da correlação entre as variáveis, podemos observar (Figura 3) que 'Breathing Problem', 'Fever', 'Dry Cough', 'Sore throat', 'Abroad travel', 'Contact with COVID Patient', e 'Attended Large Gathering' foram as variáveis com maior correlação positiva.

A alta correlação da variável 'Abroad travel' com a presença de COVID-19 é algo esperado dado a data da coleta dos dados, quando transmissão comunitária ainda não era tão comum em muitos países e ter viajado pra fora do país era um fator muito comum em casos da doença. Sendo assim, essa variável parece

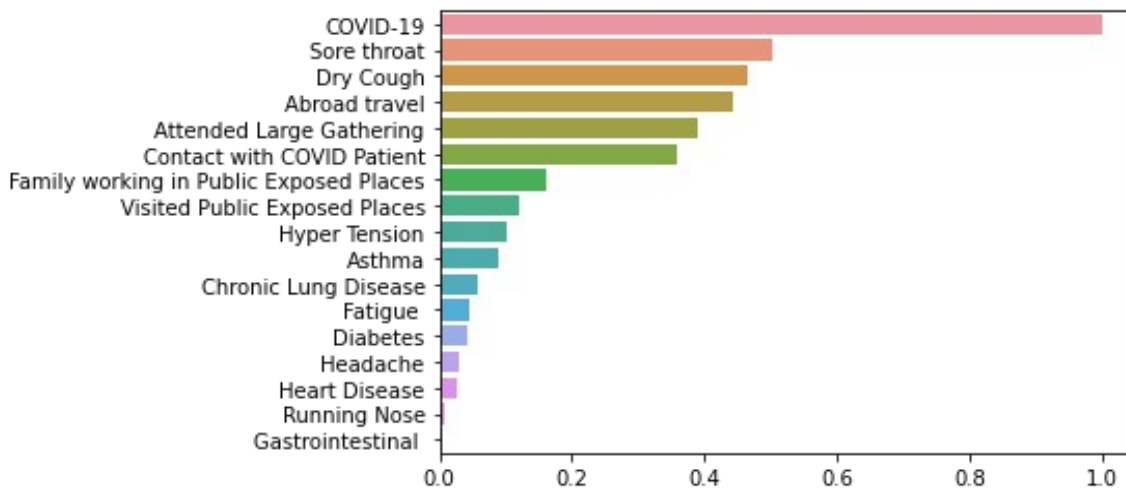
contribuir negativamente para a generalidade do modelo a ser criado, e logo será removida.

Figura 3 – Correlação entre as variáveis



Vamos observar também o módulo das correlações (Figura 4), já que variáveis com grande correlação negativa também são interessantes para a criação de um modelo de predição. Daremos preferências, em um primeiro momento, às variáveis com maior correlação com a variável alvo para a criação do modelo.

Figura 4 – Módulo correlação com a variável alvo



7. Preparação dos Dados para os Modelos de Aprendizado de Máquina

Durante as etapas anteriores as variáveis 'Wearing Masks', 'Sanitization from Market' e 'Abroad travel' foram removidas. Como pudemos observar anteriormente (Figura 2) o conjunto de dados está desbalanceado, iremos balancear os dados para a criação do modelo de classificação nos utilizando do algoritmo SMOTE para fazer uma sobreamostragem. Após a aplicação do método, a quantidade de dados em cada uma das categorias é igual (Figura 5).

Figura 5 – Conjunto balanceado

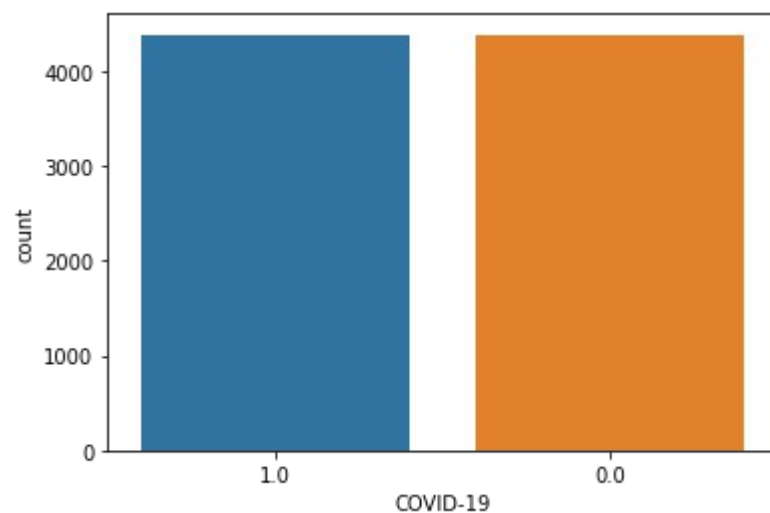


Figura 6 – Divisão do conjunto de dados

```
split = train_valid_test_split(df, target = target, train_size=0.8,  
                                valid_size=0.1, test_size=0.1)  
X_train, y_train, X_valid, y_valid, X_test, y_test = split
```

Dividimos os dados em conjuntos de treino (80%), teste (10%), e validação (10%). Agora estamos prontos para a criação do modelo de aprendizado de máquina propriamente dito. Iremos construir um modelo usando o método de regressão logística pois, entre outros motivos, torna mais fácil conseguir do modelo uma probabilidade associada à classificação (e não a classe pura e simplesmente).

8. Referências

1. Sutton, R.T., Pincock, D., Baumgart, D.C. et al. **An overview of clinical decision support systems: benefits, risks, and strategies for success.** npj Digit. Med. 3, 17 (2020). <https://doi.org/10.1038/s41746-020-0221-y>
2. Viana dos Santos Santana Í, CM da Silveira A, Sobrinho Á, Chaves e Silva L, Dias da Silva L, Santos D, Gurjão E, Perkusich A. **Classification Models for COVID-19 Test Prioritization in Brazil: Machine Learning Approach.** J Med Internet Res 2021;23(4):e27293. URL:<https://www.jmir.org/2021/4/e27293>. DOI: 10.2196/27293
3. Mirjam E Kretzschmar, Ganna Rozhnova, Martin C J Bootsma, Michiel van Boven, Janneke H H M van de Wijgert, Marc J M Bonten, **Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study.** Lancet Public Health 2020;5: e452–59, Published Online July 16, 2020. [https://doi.org/10.1016/S2468-2667\(20\)30157-2](https://doi.org/10.1016/S2468-2667(20)30157-2)
4. Hu Y, Guo J, Li G, et al **Role of efficient testing and contact tracing in mitigating the COVID-19 pandemic: a network modelling study.** BMJ Open 2021;11:e045886. doi: <https://doi.org/10.1136/bmjopen-2020-045886>.
5. Hagen, Ashley. **Supply Shortages Impacting COVID-19 and Non-COVID Testing.** American Society for Microbiology, Jan. 19, 2021. <https://asm.org/Articles/2020/September/Clinical-Microbiology-Supply-Shortage-Collecti-1>. Accessed 18/04/2020.
6. Berger, Eric. **‘There’s a lot of anxiety’: US grapples with Covid test shortage amid surge.** The Guardian, Dec. 22, 2021. <https://www.theguardian.com/us-news/2021/dec/22/us-covid-test-lines-shortages>. Accessed 18/04/2020.
7. Zoabi, Y., Deri-Rozov, S. & Shomron, N. **Machine learning-based prediction of COVID-19 diagnosis based on symptoms.** npj Digit. Med. 4, 3 (2021). <https://doi.org/10.1038/s41746-020-00372-6>