

Avaliação do Uso de IA para Auxiliar no Diagnóstico de Casos de COVID-19 para Diferentes Surtos

Felipe C. Ulrichsen¹, Alexandre C. Sena¹, Luís C. M. S. Pôrto², Karla Figueiredo¹

¹Instituto de Matemática e Estatística – Universidade do Estado do Rio de Janeiro
Rio de Janeiro – RJ – Brasil

²Laboratório de Histocompatibilidade e Criopreservação
Universidade do Estado do Rio de Janeiro – RJ – Brasil

Abstract. *Experience with COVID-19 has indicated that, in high demand situations or at underserved regions, there may be a lack of RT-PCR tests, considered standard for COVID-19 diagnosis. In this context, this work researches algorithms, based on Machine Learning (ML), for diagnosing COVID-19 only from signs and symptoms. In addition to diagnosis, this work also included studies considering outbreak waves and testing (RT-PCR and Rapid Antigen Test). The results indicate that the selected model, based on MLP Neural Networks, has a much more accurate result (70% in 1st the wave) than the one indicated by the Ministry of Health for the Rapid Test (around 25%), and indicates the changes in signs and symptoms during the evolution of the pandemic.*

Resumo. *A experiência vivida com a COVID-19 indicou que, em situações de grande demanda ou em regiões remotas e/ou carentes, pode haver falta de testes RT-PCR, considerado padrão para diagnóstico de COVID-19. Nesse contexto, este trabalho investiga algoritmos baseados em Machine Learning (ML) para diagnóstico de COVID-19, a partir de sinais e sintomas. Para além do diagnóstico, este trabalho também incluiu estudos sobre alterações nos sinais e sintomas, considerando as ondas de surtos e a qualidade dos resultados obtidos com RT-PCR e Teste Rápido. Os resultados apontam que o modelo selecionado, baseado em Redes Neurais MLP, possui resultado bem mais preciso (70%, no caso de 1ª onda) do que o indicado pelo Ministério da Saúde para o Teste Rápido (em torno de 25%), além de apontar as mudanças nos sinais e sintomas ao longo da evolução da pandemia.*

1. Introdução

A COVID-19, Doença do Coronavírus de 2019 (do inglês, *Corona Virus Disease of 2019*), que pode se manifestar em seres humanos em consequência à infecção causada pelo vírus SARS-CoV-2, foi declarada como pandemia em março de 2020 pela a OMS (Organização Mundial de Saúde).

A facilidade de transmissão [Zhu et al. 2021] [de Souza et al. 2021] e as variantes causadas pelas mutações, faz a COVID-19 comportar-se em ondas de contágio. Ainda, para realizar o controle de infectados, a OMS definiu o RT-PCR como padrão de diagnóstico para COVID-19. No entanto, em situação de escassez ou falta do teste, como ocorreu no início da pandemia, estes ficam restritos, principalmente, aos profissionais de saúde, pacientes com sintomas ou que tiveram contato com pessoas infectadas.

Na falta de testes, muitos médicos se valem dos sinais e sintomas do paciente, que são manifestações clínicas percebidas pelo paciente e por outras pessoas, fundamentais na avaliação de doenças. Tal importância torna a análise e seu estudo essenciais para melhorar a qualidade e celeridade do diagnóstico [Zhang et al. 2020] e, conseqüentemente, do tratamento.

Por outro lado, o Aprendizado de Máquina (AM) (do inglês, *Machine Learning* - ML), é uma das áreas da Inteligência Artificial (IA), que vem sendo aplicada em diversos setores, inclusive na saúde. Tendo seu uso se intensificado bastante nessa área, devido ao armazenamento digital dos dados de pacientes, hoje sua importância é reconhecida, com perspectivas inovadoras em diversas áreas da saúde [Murdoch 2013], [Wiens 2018], [Topol 2019] e [Souza et al. 2021].

Assim, *Machine Learning* possui muitas vantagens quando aplicada à área de saúde, manipulando volume alto de variáveis, de forma segura e reproduzível, em um período de tempo muito menor que um ser humano [Fang et al. 2020], [Hoffmann et al. 2020], [Yan 2020] e [Souza et al. 2021]. No entanto, muitos dos algoritmos utilizados na área de ML são considerados caixas-pretas, tornando-se uma dificuldade para a área de saúde aceitar os resultados alcançados. Nesse contexto, este trabalho tem como objetivo utilizar as técnicas de aprendizado de máquina para analisar os dados sobre *COVID-19*. Mais especificamente, deseja-se inferir o diagnóstico da *COVID-19*, a partir dos sinais e sintomas, identificar a variação dos sintomas entre as diferentes ondas de contágio, e com o apoio da técnica de *Explainable*, ajudar na análise e entendimento dos resultados obtidos por algoritmos caixa preta. Assim, é esperado que os modelos de aprendizado de máquina possam ser utilizados, por exemplo, para auxiliar profissionais da área da saúde na decisão da realização ou não do teste da *COVID-19*, reduzindo custos.

No contexto de diagnóstico de *COVID-19*, pode-se identificar um estudo realizado no Estado do Rio de Janeiro, que teve como foco a identificação de subnotificação dos casos da *COVID-19*. Ele se baseou em formulário eletrônico com preenchimentos auto-declarados de sinais e sintomas e início de sintomas *COVID-19*. Para inferir se um determinado respondente poderia estar ou não com *COVID-19*, foram desenvolvidos modelos, baseado em *Machine Learning*, para diagnóstico de *COVID-19*. O melhor modelo indicou acurácia pouco acima de 60%. A partir dos modelos construídos, foram identificados os respondentes que possivelmente estariam doentes e não foram testados, sendo considerados casos subnotificados [Souza et al. 2021].

Outro estudo [Menni et al. 2020] utilizou dados do Reino Unido e dos Estados Unidos da América. Os dados foram obtidos através de um aplicativo de celular em que os pacientes reportam os sintomas, IMC (índice de massa corpórea), sexo, doenças pré-existent, dados demográficos e o resultado do *RT-PCR*. O estudo baseou-se em dados coletados de 24/Mar/20 a 21/Abr/20. O algoritmo Regressão Logística classificou os dados, sem realizar equalização da base por resultado de *RT-PCR*, obtendo resultados com *recall* ou sensibilidade média de 65% e especificidade média de 78% para os dados do Reino Unido, e 83% e 66% para a especificidade e *recall* para os Estados Unidos. O intuito era entender se a mídia influenciou no aumento desses sintomas nos dados reportados pelos pacientes, chegando-se a conclusão de que houve um aumento, a partir do momento em que a mídia divulgou que esses sintomas correspondiam aos pacientes positivos para a *COVID-19*.

Não foram identificados, até o momento, trabalhos envolvendo análise de sinais e sintomas por ondas e tipos de testes utilizados para diagnóstico de *COVID-19*. Também não foi identificado trabalho que tenha avaliado a influência dos sinais e sintomas informados no diagnóstico. Outro aspecto importante na diferenciação dos demais estudos é utilizar-se apenas de sinais e sintomas descritos pelos profissionais de saúde e pacientes como atributos.

O restante deste documento está dividido em mais 4 seções. A Seção 2 faz um resumo dos Fundamentos Teóricos. Em seguida, a Seção 3 apresenta a metodologia utilizada neste trabalho para realizar a avaliação dos sinais e sintomas em diferentes períodos e testes utilizados durante a pandemia de *COVID-19*. A Seção 4 descreve o Estudo de Caso e, por fim, a Seção 5 conclui e apresenta perspectivas de trabalhos futuros.

2. Fundamentação Teórica

Nesta seção, são apresentados sucintamente os algoritmos utilizados neste trabalho.

2.1. *Random Forrest*

O algoritmo Floresta Aleatória (*Random Forest - RF*) é um algoritmo do tipo aprendizado de comitês, método que gera muitos classificadores e combina os seus resultados. O algoritmo usa *bootstrap aggregating* ou *bagging*, selecionando registros aleatoriamente com reposição para serem usados na construção de cada árvore, reduzindo a variância sem prejudicar o bias [Breiman 2001]. Breiman [Breiman 1996] também observou que, em um conjunto de árvores de decisão havia correlação profunda entre elas, e uma forma de reduzir esse efeito é a seleção aleatória de atributos que podem estar disponíveis para a construção de cada árvore, diminuindo a correlação entre as árvores de decisão.

2.2. *Multi-Layer Perceptron*

O algoritmo *Multi-Layer Perceptron* é um algoritmo inspirado em neurônios biológicos. Ele é baseado na construção de redes de neurônios artificiais (interconectando os neurônios entre as camadas), para aprender a mapear relações de dados de entradas em dados de saída em problemas supervisionados, por meio do ajuste dos pesos sinápticos, baseados nos erros identificados durante o processo de aprendizado. A correção, ou seja, os ajustes sinápticos são realizados por retropropagação dos erros, que é obtido a cada ciclo de apresentação (*batch*) de dados ao modelo. Este algoritmo é conhecido como um aproximador universal [Rumelhart 1986].

2.3. Regressão Logística

A regressão logística é uma técnica estatística simples, que se tornou largamente utilizada em muitas áreas. Ela tem por objetivo gerar, a partir de dados de entrada (sejam eles numéricos ou nominais), um modelo linear que permita a predição de valores definidos por uma variável categórica, em geral binária [Brant 1996].

2.4. *Shapley Additive Explanations*

Os autores do trabalho [Lundberg and Lee 2017] incluíram o Explicações Aditivas de Shapley (do inglês *Shapley Additive Explanations*) como uma abordagem para explicar o resultado de algoritmos de aprendizagem de máquinas, viabilizando a interpretação de modelos. A essência do valor de SHAP é medir a contribuição de cada variável para o resultado final .

3. Metodologia

Uma visão geral de todo o processo é apresentada na Figura 1. A metodologia pode ser dividida em três etapas distintas: aquisição de dados, pré-processamento e aplicação e análise dos resultados de classificação obtidos por *Machine Learning*.

A etapa de aquisição dos sinais e sintomas e dos resultados de testes, Figura 1.(2), é feita a partir de sistemas de saúde (Figura 1.(1)). Por sua vez, a etapa de pré-processamento, Figura 1.(3), é iniciada com a junção dos dados correspondentes identificados nas bases de dados (sinais e sintomas + resultados de testes). É importante destacar que os dados não precisam ser identificados, ou seja, não é necessário que os dados pessoais ou sensíveis estejam disponíveis. Em seguida, é realizada a normalização dos dados de natureza quantitativa, Figura 1.(4), quando for o caso. Os atributos qualitativos, ordenáveis ou não, devem ser transformados em dados numéricos. No caso deste trabalho, os atributos são categóricos e dicotômicos, sendo portanto transformados em zero (0) ou um (1). Ainda na etapa de pré-processamento, os dados são separados por ondas de contágio, Figura 1.(5), isto é, os dados são divididos com base nas datas de início e fim dos surtos, e têm suas bases balanceadas com respeito ao atributo de saída. O próximo passo é a seleção de atributos, Figura 1.(6), com a aplicação de técnicas do tipo *filter*, selecionando variáveis mais relevantes para os algoritmos de *Machine Learning* (ML), visando maximizar a acurácia da classificação.

Com os dados pré-processados, a etapa 3 é iniciada com o treinamento dos algo-

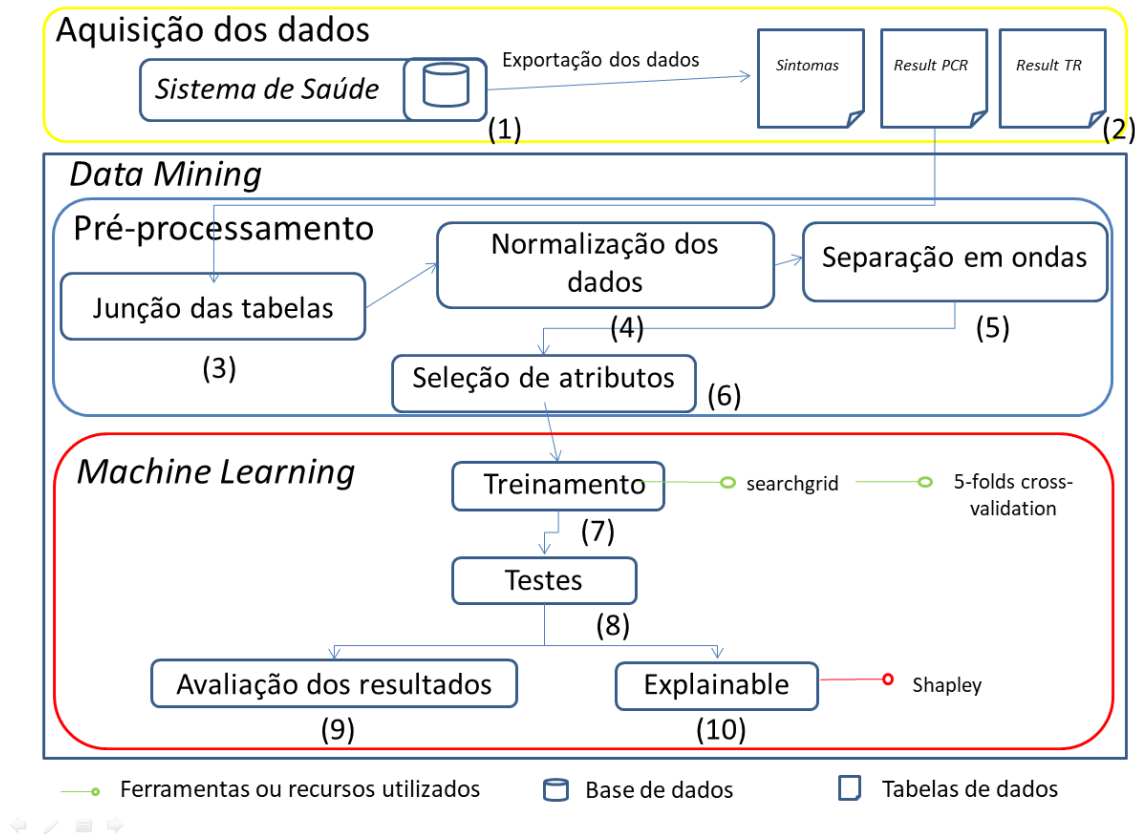


Figura 1. Diagrama Simplificado da Metodologia

ritmos de *Machine Learning*, Figura 1.(7). Cada modelo é proposto a partir da definição de seus hiperparâmetros, seguida pela avaliação de seus resultados por validação cruzada, que pode se basear em uma ou mais métricas (em geral, as médias da acurácia, recall, precision e f1 e a curva ROC/AUC) obtidas com os n modelos durante a validação cruzada. Os melhores resultados médios de validação cruzada identificam os algoritmos/modelos mais adequados.

Uma vez selecionados os modelos, é inicializada a fase de testes(Figura 1.(8)), que é avaliada (Figura 1.(9)) pelas mesmas métricas usadas na validação e, em geral, acrescidas da matriz de confusão. Por fim, para se entender a importância de cada atributo no resultado obtido, é realizada a fase compreensão dos resultados (*Explainable*), Figura 1.(10), utilizando, por exemplo, o método de *Shapley*.

4. Resultados

Nesta seção serão apresentados e avaliados os resultados considerando a investigação de diagnóstico, sob a ótica das diferentes ondas do surto de *COVID-19*, além da qualificação dos testes empregados.

4.1. Pré-Processamento

O conjunto de dados disponíveis da Policlínica Piquet Carneiro (PPC) da UERJ, três tabelas de fontes distintas, foram unificadas. A primeira tabela contém a data do início dos sinais e sintomas de pacientes que procuraram a clínica com suspeita de *COVID-19*, além de todos os sinais e sintomas relatados pelo paciente em campo textual livre. As outras duas tabelas possuem os resultados dos exames *RT-PCR* e Teste Rápido (TR) desses pacientes.

Os sintomas descritos, de forma livre em formulário, pelos pacientes foram unificados (por exemplo, “cefaléia” e “dor de cabeça”), a partir da análise de três avaliadores.

Em seguida, cada um dos sintomas indicados foram definidos de forma dicotômica (zero e um) e separados como atributos distintos. Assim, um determinado sintoma foi definido como “1” para o paciente que indicou que teve o sintoma e “0” caso contrário. A coluna diagnóstico também foi categorizada como “1” caso o resultado do teste (*RT-PCR* ou TR) tenha sido positivo e “0” quando negativo para *COVID-19*.

Foi feito um corte para se considerar os 19 sintomas mais frequentes na base de dados, após a avaliação de especialistas da área da saúde. O objetivo é, em um primeiro momento, considerar de forma mais ampla os sinais e sintomas relatados, mesmo que tenham baixa representatividade na base de dados, e deixar que a etapa de seleção de variáveis e os algoritmos de ML expurguem os sinais e sintomas que não ajudem a discriminar o diagnóstico.

A Figura 2 apresenta a prevalência dos 19 sinais e sintomas em paciente com resultados negativos e positivos, segundo o TR (Figura 2-parte superior) e o teste *RT-PCR* (Figura 2-parte inferior). A Tabela 1 complementa a Figura 2 com o quantitativo de registros de pacientes com sinais e sintomas relacionados ao tipo de teste e seus resultados. Destaca-se o maior número de teste *RT-PCR* e o desbalanceamento das bases com relação aos resultados dos testes.

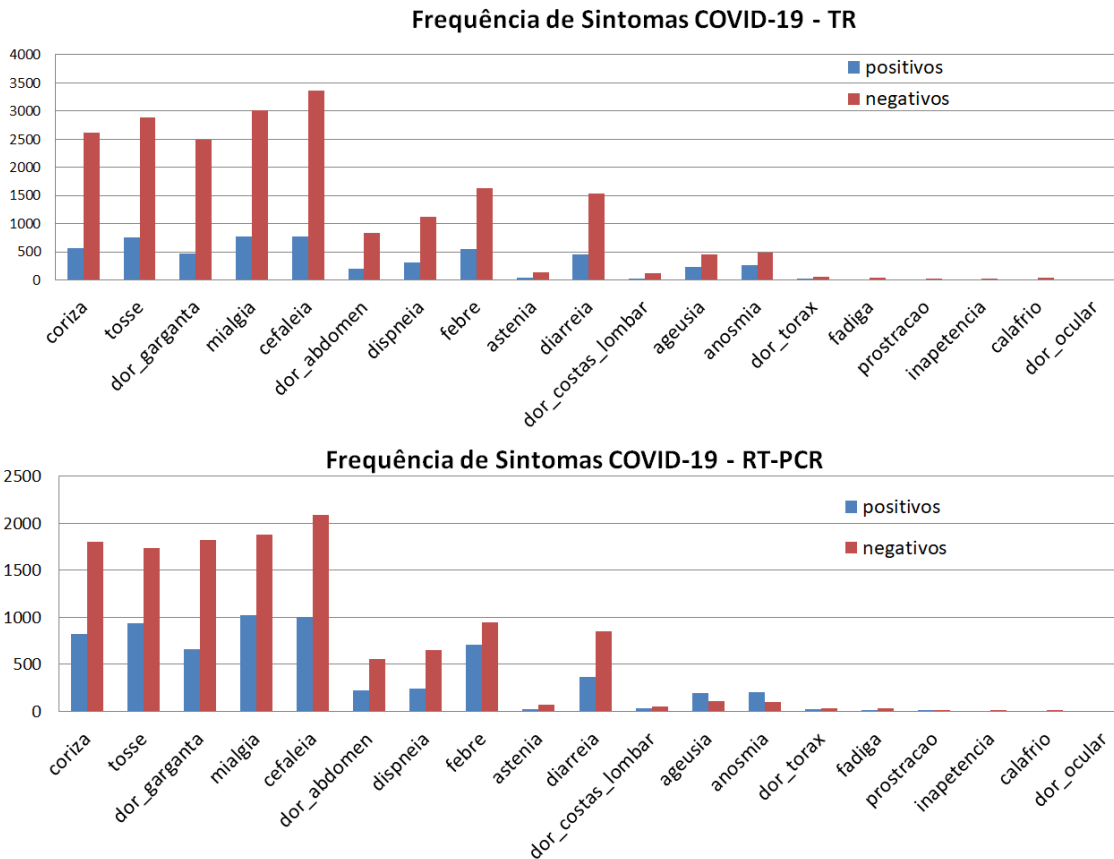


Figura 2. Gráficos Frequência de Sinais e Sintomas - RT-PCR e Teste Rápido (TR)

Tabela 1. Total de registros de pacientes por tipo de teste

	TR tot	TR pos	TR neg	RT-PCR tot	RT-PCR pos	RT-PCR neg
#registros	6200	1169	5031	4858	1587	3271

A Figura 3 tem por objetivo a identificação das ondas de surto da *COVID-19*, considerando o gráfico de média móvel de sete dias dos casos confirmados da *PPC* (UERJ), pareado ao gráfico de média móvel dos casos notificados da cidade do Rio de Janeiro. Nela é possível identificar que a primeira e terceira ondas no gráfico do Rio de Janeiro coincidem com a primeira e terceira ondas da *PPC* (UERJ), indicando que o início da primeira onda se deu no dia 18/03/2020 e o final no dia 18/06/2020, e 18/10/2020 e 18/02/2021 foram definidos como início e fim da segunda onda, respectivamente

Em seguida, o conteúdo dessa base foi separado segundo as ondas de contágio de *COVID-19* e as bases foram balanceadas. Com as datas definidas, pôde-se separar as bases em treinamento e teste para avaliação dos algoritmos de ML, considerando os dados pertinentes a 1ª e 2ª ondas e tipos de testes realizados (ver Tabela 2).

O último passo na etapa de pré-processamento foi a aplicação de dois métodos de seleção de variáveis: *ReliefF* [Kononenko 1997] e Seleção de Atributos Baseada em Correlação (*CFS*) (do inglês, *Correlation-based Feature Selection*) [Mark A. Hall 1999]. Destaca-se que as bases foram separadas de forma aleatória em *hold-out*, com 10% para

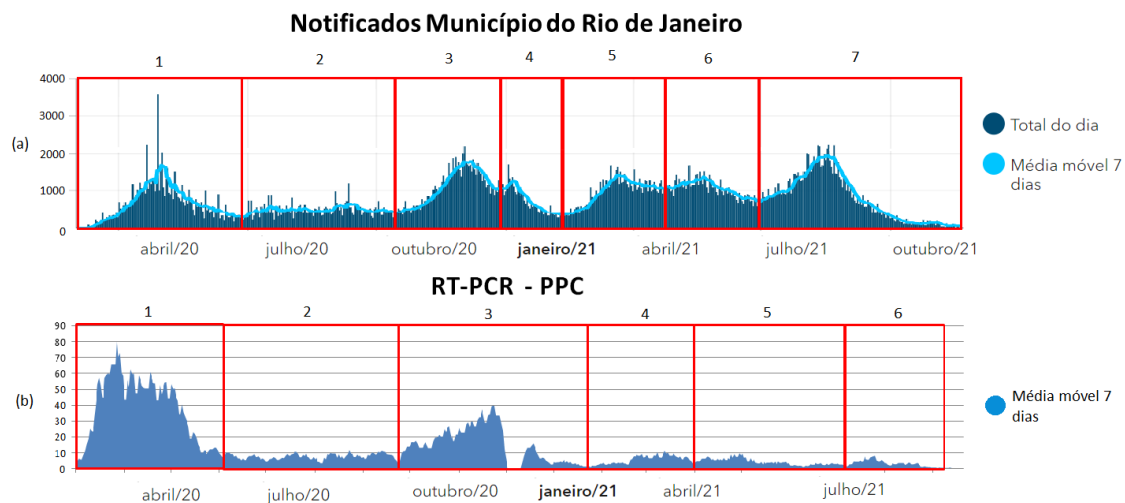


Figura 3. Gráfico das ondas do surto de COVID-19

Tabela 2. Quantidade de registros por base de dados: aproximadamente 90% para treinamento e validação e 10% para teste

	PCR tot	PCR 1a	PCR 2a	TR tot	TR 1a	TR 2a	PCR+TR tot	PCR+TR 1a	PCR+TR 2a
Trein. e Val.	2436	1208	702	1990	1748	67	4426	2926	769
Teste	272	136	78	222	196	9	494	332	87

todas as bases devido a quantidade pequena de dados para algumas bases, sendo avaliadas por essas técnicas, resultando em novas bases de dados com conjunto de atributos distintos, conforme Tabela 3.

Além dos subconjuntos de atributos indicados em cada teste/onda por ReliefF e CFS, também foram consideradas as agregações de atributos indicados por ambas as técnicas para cada base. Assim, além das nove bases envolvendo os 19 atributos, também foram avaliadas as 18 bases obtidas com a seleção de variáveis na Tabela 3 e mais as nove bases formadas pela conjugação de atributos selecionados pelas duas técnicas de seleção de variáveis para cada uma das 9 bases, totalizando 36 diferentes conjuntos de dados.

Pode-se observar, na Tabela 3, que na 1ª e 2ª ondas, tanto a ageusia como a anosmia estavam entre os atributos selecionados por ambas as técnicas.

Tabela 3. Seleção de Variáveis (Sinais e Sintomas) por base de teste e onda

	RT-PCR		TR		RT-PCR+TR	
	ReliefF	CFS	ReliefF	CFS	ReliefF	CFS
Total	9,16,1,8	19,11,17 15,10,1,2	9,18,2,4 16,14,8,11	11,15,3,7,1,2 13,18,16,12	19,11,17,15 10,1,2,13,16	9,16,15,1,2,19 11,17,10,13
1a Onda	8,15,16,1 9,4	15,3,10,1 2	17,19,15,1 14,2,11,18	11,15,3,7,1,2 13,18,16,12	2,19,1,16,4 17,9,15,8	11,15,3,1,2,13 16,4,12
2a Onda	9,1,2,19 15,12,8	19,9,15,1 2,3,12	5,19,2,17 7,9,8,10,3	19,17,5,3,10 1,2	9,1,2,19,15 12,11,8	19,9,15,3,1,2 13,16,12

*ageusia (1), anosmia (2), astenia (3), calafrio (4), cefaleia (5), coriza (6), diarreia (7), dispneia (8), dor_abdomen (9), dor_costas_lombar (10), dor_garganta (11), dor_ocular (12), dor_torax (13), fadiga (14), febre (15), inapetência (16), mialgia (17), prostração (19) e tosse (19)

4.2. Classificando Sinais e Sintomas por Machine Learning para o Diagnóstico de COVID-19

Os dados utilizados nesse trabalho foram obtidos de pacientes que procuraram a Policlínica Piquet Carneiro (PPC) no período de março de 2020 a outubro de 2021 e foram testados após terem relatado seus sinais e sintomas. Conforme já foi mencionado na Seção 4.1, as bases exploradas pelos algoritmos foram balanceadas, mantendo a distribuição de registros entre treino/validação e teste indicada na Tabela 2. A base relativa ao TR 2ª onda foi considerada inadequada para o treinamento dos modelos devido ao reduzido número de registros, mas estes foram incluídos nas bases considerando a junção dos testes *RT-PCR* e TR totais e separados em 1ª e 2ª onda.

4.2.1. Configurações Treinamento dos Algoritmos

Para realizar a análise, a partir do diagnóstico foram usados três algoritmos: *Random Forrest-RF*, *Multi-Layer Perseptron-MLP* e Regressão Logística-RL, com diversas parametrizações, com o objetivo de buscar a melhor acurácia para o diagnóstico, para cada uma das 36 bases apresentadas na Subseção 4.1. Assim, as Tabelas 4 e 5 indicam os valores dos hiperparâmetros explorados nos algoritmos (usando a função *searchgrid* do *Python*), combinados exaustivamente.

Como critério de avaliação foi utilizado *cross-validation* com $k\text{-fold} = 5$ em todos os algoritmos/bases avaliados. Assim, para cada algoritmo com

Para a RL a seleção de variáveis proposta pelo RelifF reduziu, de forma geral, as médias das métricas e o CFS e a soma das duas técnicas manteve. Isso aponta que o conjunto de variáveis indicadas pelo CFS é adequado para a RL. Para o caso do *RF*, todos os modelos criados a partir de bases com variáveis selecionadas pelas duas técnicas, quando comparadas a totalidade dos atributos, mostraram menores métricas. Para os

Tabela 4. Configurações de parâmetros usado no SearchGrid para *Random Forrest* em todas as configurações de bases com 100 estimadores

<i>Random Forrest</i>	
max_depth	3, 4, 5, 6, 7, 8, 9, 10
min_samples_split	2, 4, 8, 12, 16
criterion	gini, entropy
max_features	auto, 2, 3, 4, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
min_samples_leaf	2, 3, 4, 5, 6, 7, 8

Tabela 5. Configurações de parâmetros usado no SearchGrid para *Multi-Layer Perseptron* e Regressão Logística em todas as configurações de bases

<i>Multi-Layer Perseptron</i>		Regressão Logística	
n. neurônios	[5, 120]	tol	0.5, 0.001
learning_rate	adaptive, constant	C	1.0, 2.0, 3.0
learning_rate_init	0.06 e 0.001	fit_intercept	True, False
nesterovs_momentum	False e True	solver	lbfgs, saga
solver	sgd e adam		

Tabela 6. Validação com acurácia média (k-fold=5) com o melhor conjunto de hiperparâmetros avaliado - Base com todos os 19 atributos

	MLP	RF	RL
<i>RT-PCR</i> Total	0.70	0.67	0.65
<i>RT-PCR</i> 1ª onda	0.64	0.62	0.61
<i>RT-PCR</i> 2ª onda	0.80	0.70	0.68
TR Total	0.70	0.64	0.62
TR 1ª onda	0.73	0.63	0.61
<i>RT-PCR</i> +TR Total	0.67	0.65	0.63
<i>RT-PCR</i> +TR 1ª onda	0.68	0.62	0.61
<i>RT-PCR</i> +TR 2ª onda	0.78	0.69	0.66

modelos baseados em *MLP* a diferença das métricas sem e com seleção de variáveis é ainda maior. Por esse motivo e pelo reduzido espaço, não serão apresentados os resultados produzidos com as bases que consideraram seleção de variáveis.

A Tabela 6 apresenta os melhores resultados médios de acurácia de validação (considerando todas as combinações de hiperparâmetros indicada), para todos os algoritmos avaliados. Apesar da métrica escolhida para a seleção do melhor modelo ter sido a acurácia, as métricas *precision*, *recall*, *f1* e área sobre a curva ROC, também foram calculadas (mas foram omitidas por motivo de espaço) e serviram, nessa ordem, para desempate em caso de valores de acurácia muito próximos.

Conforme observado na Tabela 6, algoritmo *MLP* supera o *RF* e a *RL* em todas as bases avaliadas. O mesmo ocorreu com os resultados para as bases com atributos indicados na Tabela 3, para cada uma das bases da avaliadas (Tabela 2). Assim, devido ao espaço reduzido, serão apresentados na Tabela 7 apenas os valores dos hiperparâmetros dos modelos de redes *MLP* para cada base investigada, considerando os 19 atributos, pois este algoritmo alcançou os melhores resultados (Tabela 6). Em todos os modelos a taxa de *momentum* foi igual a 0,9, o número de épocas foi fixado em 50 e a paciência fixada em 5.

Aplicando-se o teste *Mann-Whitney* obtem-se p-valores < 0.05 para a comparação dos resultados obtidos pela *MLP* e *RF* (p-valor= 0,009) e pela *MLP* e *RL* (p-valor= 0,002), mas obtém-se resultado p-valor $> 0,05$ (p-valor= 0,110) para *RF* e *RL*. Isso significa que os resultados da rede neural realmente são significativamente menores e podem ser considerados melhores do que os obtidos pelo *RF* e *RL*, o mesmo não pode ser dito para os resultados do *RF* e *RL*, ou seja, são resultados equivalentes.

Pode-se destacar que a avaliação da base *RT-PCR* 2ª onda, de forma geral, apresentou o melhor desempenho para o diagnóstico, refletindo também no *RT-PCR* Total. Isso pode indicar que estes sinais e sintomas foram mais determinantes em discriminar quais pacientes tinham ou não *COVID-19*, atrelado ao fato do teste ser mais confiável nos resultados gerados, ou seja, os rótulos positivos ou negativos mais confiáveis. Por outro lado, devido às falhas, já relatadas do TR ¹, quanto à qualidade dos resultados, pode-se acreditar que isto deve ter ajudado a reduzir o desempenho dos modelos para as bases associadas a este teste.

¹<https://www.sbac.org.br/blog/2020/04/02/teste-rapido-da-covid-19-falsos-negativos/>

Tabela 7. Melhores hiperparâmetros por base da dados - *Multi-Layer Perceptron* - Todos os 19 atributos

	learning_rate	learning_rate_init	nesterovs_momentum	solver
<i>RT-PCR</i> Total	adaptive	0.06	False	sgd
<i>RT-PCR</i> 1ª onda	adaptive	0.06	False	sgd
<i>RT-PCR</i> 2ª onda	adaptive	0.001	True	adam
TR Total	adaptive	0.001	True	adam
TR 1ª onda	adaptive	0.001	True	adam
TR 2ª onda	adaptive	0.06	False	adam
<i>PCR+TR</i> Total	constant	0.0001	True	sgd
<i>PCR+TR</i> 1ª onda	adaptive	0.001	True	adam
<i>PCR+TR</i> 2ª onda	adaptive	0.001	True	adam

os hiperparâmetros momentum = 0.9, n_iter_no_change = 5 e max_iter = 50 não variaram

Tabela 8. Métricas para os conjuntos de teste - *Multi-Layer Perceptron* - Bases com todos os 19 atributos

	Acurácia	<i>Precision</i>	<i>Recall</i>	<i>f1</i>
<i>RT-PCR</i> Total	0.64	0.71	0.64	0.64
<i>RT-PCR</i> 1ª onda	0.70	0.84	0.70	0.70
<i>RT-PCR</i> 2ª onda	0.62	0.56	0.62	0.62
TR Total	0.51	0.47	0.51	0.51
TR 1ª onda	0.59	0.60	0.59	0.59
<i>RT-PCR+TR</i> Total	0.56	0.58	0.59	0.59
<i>RT-PCR+TR</i> 1ª onda	0.63	0.64	0.63	0.63
<i>RT-PCR+TR</i> 2ª onda	0.60	0.53	0.60	0.61

4.2.2. Avaliação dos Resultados de Teste

Observando a Tabela 8, percebe-se que o modelo baseado em redes neurais alcançou melhores métricas: 70% na acurácia e em *f1* usando a base de dados *RT-PCR* 1ª onda. Chama a atenção também os 84% da métrica *precision*, como pode ser observado na Tabela 8. A *precision* alta indica que quando o modelo indica que o paciente tem *COVID-19*, a partir dos sinais e sintomas, acerta 84% das vezes. Os resultados com modelos utilizando as base de dados do TR confirmaram os piores desempenhos em todos algoritmos, seguindo o que os resultados de validação (Tabela 7) já indicavam. No entanto, ressalta-se a inversão de melhor resultado para o *RT-PCR* na 1ª onda.

Com o objetivo de agregar explicações para os resultados obtidos, mas devido ao reduzido espaço disponível, serão apresentados na Figura 4, apenas os gráficos de valores SHAP dos preditores de *COVID-19* para o modelo *MLP* para as bases *RT-PCR* 1ª e 2ª onda, que foram os que mais se destacaram nos resultados.

O gráfico é criado por todos os pontos dos dados de treinamento de cada base. Como pode-se observar, a febre, mialgia e anosmia são as três variáveis mais importantes na 1ª onda. No entanto, especificamente ter anosmia ou ageusia indicou impacto para classificar *COVID-19*, e não ter, uma ou outra, não contribuiu significativamente para a qualidade da classificação. Além disso, não ter coriza ou a soma dos outros 10 atributos,

de forma generalizada, apontou impacto para a classificação da *COVID-19*. Ressalta-se que não ter mialgia alcançou impactos maiores na indicação de ausência de *COVID-19*.

Já na 2ª onda, a febre, tosse e dor de garganta tiveram maior impacto nos resultados de classificação. Destacando-se que ter febre ou não ter dor de garganta ajudou na identificação da *COVID-19*, e ter dispneia ou dor abdominal contribuiu para indicar ausência de *COVID-19*. Dessa forma, pode-se constatar que, para os pacientes avaliados por *RT-PCR* na PPC na 1ª e 2ª ondas, houve algumas diferenças nos sinais e sintomas que poderiam ajudar no diagnóstico para *COVID-19*. De forma geral, os sinais e sintomas, que estão presentes em ambas os gráficos, concordam com relação às contribuições ou não, a exceção da coriza que aparece invertido. Além disso, destaca-se a anosmia e ageusia entre os 10 sinais e sintomas da primeira onda e a sua ausência na segunda onda.

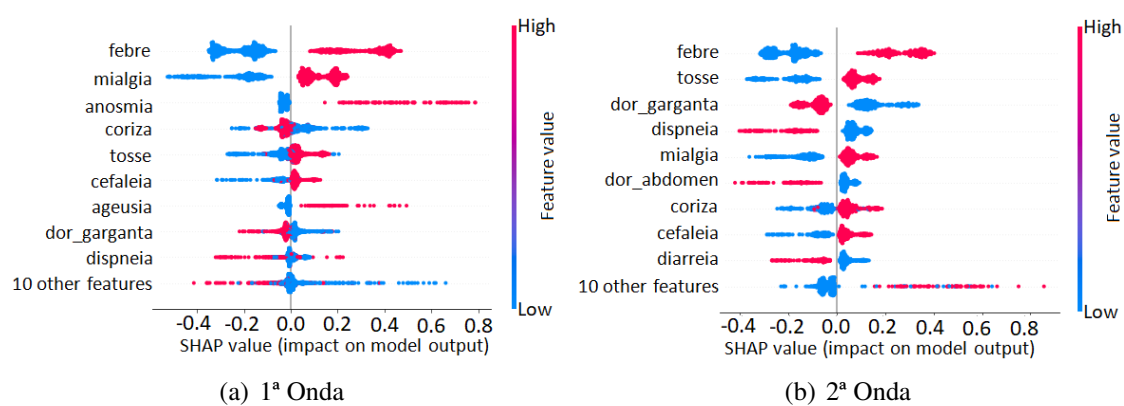


Figura 4. Gráfico SHAP com os impactos dos atributos na qualidade da classificação - MLP - RT-PCR - 19 atributos

5. Conclusão

Esta metodologia mostrou-se útil no estudo e avaliação da prevalência de sinais e sintomas da *COVID-19*, especialmente em função de novas variantes, que podem ocorrer ao longo do tempo, servindo como uma importante ferramenta para a área de saúde. Por meio desse estudo também foi possível identificar as diferenças entre os dois tipos de testes: *RT-PCR* e TR e entre as onda de surto.

Os modelos de ML propostos neste trabalho foram capazes de aprender a indicar diagnósticos com base em sinais e sintomas, melhor do que os TRs aplicados em pacientes, que mostraram erros de até 75%, tornando-se alternativas para apoio ao diagnóstico, principalmente em ambientes remotos ou com carência de recursos. Destaca-se a performance dos modelos MLP obtidos com as bases de pacientes que foram testados pelo *RT-PCR*. A análise com método *Explainable* também se mostrou útil ao constatar e apontar as diferenças nos sinais e sintomas entre 1ª e 2ª ondas. Como possibilidades de extensão deste trabalho, deseja-se restringir a coleta para testes de pacientes, apenas para os intervalos de maior acurácia dos exames *RT-PCR* e Teste Rápido, além de estudo e análise de sinais e sintomas que correspondem a diferentes diagnósticos.

Agradecimentos

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Financiamento 001 e Projeto 88881.506840/2020-01.

Referências

- Brant, R. (1996). Digesting logistic regression results. *The American Statistician*, 50(2):117–119.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- de Souza, U., dos Santos, R., Campos, F., Lourenço, K., da Fonseca, F., and Spilki, F. (2021). High rate of mutational events in sars-cov-2 genomes across brazilian geographical regions. *Viruses*, 13.
- Fang, Y., Zhang, H., J., X., Lin, M., Ying, L., Pang, P., and Ji, W. (2020). Sensitivity of chest ct for covid-19: Comparison to rt-pcr. *Radiology*, 296:115–117.
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Kruger, N., Herrler, T., and Erichsen, S. (2020). Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*, 16:271–280.
- Kononenko, I. (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence*, 7(1):39–55.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proc. of the 31st Inter. Conference on Neural Information Processing Systems, NIPS' 17*, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.
- Mark A. Hall, L. A. S. (1999). Feature selection for machine learning: Comparing a correlation based filter approach to the wrapper. *Proceedings of the Twelfth International FLAIRS Conference*, 30:4765–4774.
- Menni, C. et al. (2020). Real-time tracking of self-reported symptoms to predict potential covid-19. *Nature Medicine*, 26(7):1037–1040.
- Murdoch, T.B., D. A. (2013). The inevitable application of big data to health care. *Jama*, 309(13)::1351–1352.
- Rumelhart, D. J. M. (1986). Parallel distributed processing: explorations in the microstructure of cognition. *Cambridge: MIT Press*.
- Souza, M., Figueiredo, K., Porto, L., and Medronho, R. (2021). *Experiências e impacto da pandemia pela Covid-19 no complexo de Saúde UERJ*, volume 1. Rio de Janeiro.
- Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. *Basic Books*, 1.
- Wiens, J., S. E. (2018). Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66:149–153.
- Yan, L. (2020). A machine learning-based model for survival prediction in patients with severe covid-19 infection. *medRxiv*.
- Zhang, X. et al. (2020). Biological, clinical and epidemiological features of covid-19, sars and mers and autodock simulation of ACE2. *Infect Dis Poverty*, 20):99.
- Zhu, M. et al. (2021). Molecular phylogenesis and spatiotemporal spread of sars-cov-2 in southeast asia. *Public Health*, 9.