XXXI Congresso de Iniciação Científica Unicamp 2023

# Uma investigação sobre métodos de interpretabilidade local para predições de diagnósticos médicos com base em aprendizado de máquina

**Autores:**
**Lucas Tramonte - UNICAMP**
**Prof. Cristiano Torezzan(Orientador) - UNICAMP**

## 1 Introdução

Modelos de Aprendizado de Máquina (AM) têm sido amplamente utilizados nas mais diversas áreas, impulsionando importantes mudanças de paradigma em processos decisórios que, até recentemente, eram essencialmente feitos por seres humanos. Na área de saúde, por exemplo, modelos de AM têm sido utilizados para o auxílio de profissionais em predições e diagnósticos, com o enfoque na interpretabilidade dos mesmos. Dentre os principais modelos utilizados, destaca-se a Regressão Logística com o uso do *odds ratio* (OR), o qual mede a relação entre a chance de ocorrência de um evento quando uma determinada variável aumenta em um fator de 1, o que produz resultados facilmente interpretáveis, configurando um modelo preditivo útil para a área da saúde [4].

Por um outro lado, para o uso dos demais modelos como as redes neurais artificiais ou XG-Boost, métodos de interpretabilidade a posteriori como SHAP e LIME podem ser utilizados para adicionar uma camada de significados que permita ao profissional de saúde identificar, por exemplo, qual a importância dos principais atributos para a classificação. O método SHAP, um dos mais utilizados para esta finalidade, propõe explicar as predições locais de um determinado modelo, ao realizar aleatoriamente diversas coalizões de atributos e determinar os seus respectivos efeitos no mesmo, enquanto o método LIME busca obter interpretabilidade a posteriori por meio da aproximação local da função (*black-box*) por uma função mais simples e interpretável. Nesse sentido, o presente projeto teve como objetivo realizar um estudo teórico-conceitual sobre os métodos de interpretabilidade SHAP e LIME, bem como analisar o desempenho desses métodos em aplicações na área da saúde, com um enfoque na predição do patógeno SARS-CoV-2 a partir de sintomas. Para a aplicação prática, utilizamos dados reais sobre Covid-19 disponibilizados em [1] e implementamos os modelos de RL, XG-boost e RF na linguagem python, com apoio de bibliotecas como Sckit Learning [7]. Os resultados obtidos com o método SHAP foram comparados com os OR de modelos de regressão logística implementados neste estudo e também com resultados publicados em [1], onde os autores utilizam métodos de AM para predizer resultados de testes do SARS-CoV-2 com base em sintomas relatados e características sócio-demográficas, mas não exploram aspectos mais avançados de interpretabilidade.

# 2 Metodologia

## 2.1 Regressão Linear

Modelos lineares são amplamente utilizados na área da saúde, uma vez que são passíveis de interpretação a partir dos pesos de cada atributo, possibilitando determinar a contribuição $\phi_j$ de cada atributo em uma predição[4]:

$$\sum_{j=1}^{p} \phi_j(f) = f(x) - E(f(x)), \tag{1}$$

sendo $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ a predição para a instância x e $E(f(x))$ o valor predito médio para esta mesma instância.

### 2.1.1 Regressão Logística

Dentre os principais modelos lineares utilizados, destaca-se a Regressão Logística, na qual a variável dependente Y é binária e há um conjunto de $p$ variáveis independentes:

$$P(y^{(i)} = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)})}}, \tag{2}$$

sendo $\beta_0$, $\beta_1$, ..., e $\beta_p$ os coeficientes do modelos determinados pelo método da máxima verossimilhança.

Manipulando a equação (1) , obtém-se os OR, os quais medem a chance da ocorrência da variável dependente a partir de um determinado atributo, comparado com a chance desta mesma ocorrência na ausência do atributo em questão. Desta forma, os OR produzem resultados facilmente interpretáveis, configurando um modelo preditivo útil para a área da saúde [4]:

$$\begin{aligned} \text{OR} = \frac{P(y^{(i)} = 1)}{1 - P(y^{(i)} = 1)} \\ = e^{-(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)})} \end{aligned} \tag{3}$$

Neste trabalho, o OR da regressão logística é utilizado como referência de comparação (*baseline*) nos experimentos numéricos. A ideia principal é comparar os valores obtidos com os métodos SHAP e LIME com os OR da regressão logística.

## 2.2 Shapley Values

A fundamentação teórica do método SHAP está baseada na teoria econômica dos jogos. A ideia central é atribuir custos marginais para cada atributo e analisar sua contribuição média sobre todas as combinações de coalizões possíveis desses atributos. O valor médio da contribuição de uma dada coalizão denominado Shapley value é definido por [5]:

$$\begin{aligned} \phi_j(val) = \sum_{S \subseteq \{1,\ldots,p\} \setminus \{j\}} \frac{|S|!(p - |S| - 1)!}{p!} \\ \cdot (val_x(S \cup \{j\}) - val_x(S)) \end{aligned} \tag{4}$$

sendo $S$ o subconjunto dos atributos utilizados no modelo, $x$ o vetor de valores dos atributos da instância a ser explicada e $val_x$ a predição dos atributos do conjunto S que foram marginalizados sobre os atributos que não pertencem à este conjunto. O valor de S é dado por:

$$val_x(S) = \int f(x_1, \ldots, x_p) dP_{x \notin S} - E_X[f(X)]. \tag{5}$$

## 2.3 LIME

O método LIME consiste em um modelo interpretável $g$ que busca explicar predições individuais, aproximando-as do modelo original $f$ que contém o efeito black-box. Nesse sentido, o LIME visa compreender o comportamento deste modelo gerando um novo conjunto de dados, o qual contém as amostras originais e as amostras que sofreram uma determinada perturbação, permitindo observar a proximidade entre tais instâncias de forma ponderada a partir do treinamento dos dados com o modelo $g$:

$$\underset{g \in G}{\text{argmin}} L(f, g, \pi_x) + \Omega(g) = \text{explanation}(x), \tag{6}$$

onde $\Omega(g)$ é a complexidade do modelo original, $G$ a família de todos os possíveis modelos interpretáveis e L uma função de perda (*loss function*) que deve ser minimizada:

$$L(f, g, \pi_x) = \sum_{z, z' \in Z} \pi_x(z) \cdot [f(z) - g(z')]^2, \tag{7}$$

sendo $(z', f(z))$ o conjunto de dados das amostras pertubadas e $\pi_x(z)$ a medida de proximidade que determina o tamanho da vizinhança em torno da instância $x$ [6].

## 2.4 Materiais

Os principais materiais de apoio utilizados para o desenvolvimento desse projeto foram artigos científicos que abordam aplicações de métodos de aprendizado de máquina para problemas da área de saúde (ex. [3], [6], [2]).

No âmbito metodológico, os testes foram realizados por meio de implementações computacionais em linguagem de programação Python, com apoio de bibliotecas específicas de aprendizado de máquina como Scikit-learn[7] e Imodels [8].

# 3 Resultados e Discussão

Nos testes realizados com base nos dados disponibilizados em [1], busca-se compreender a relação entre os sintomas Tosse, Febre, Dor de Garganta, Coriza, Mialgia, Enjoo, Diarreia, Perda de Olfato e Falta de Ar, com o resultado binário do teste do COVID 19. As principais informações estatísticas sobre a amostra populacional em questão estão expostas na Figura 1 , na qual foram apresentados os valores percentuais para as variáveis categóricas, e a mediana para as variáveis contínuas. Com isso, torna-se possível perceber um desbalanceamento na variável de resposta, com apenas 9101 (14.25) dos indivíduos diagnosticados com o patógeno SARS-COV-2. Dessa forma, os modelos de Regressão Logística, XG-boost e Random Forest foram aplicados no conjunto de dados, com o uso de uma técnica de sub-amostragem (*undersampling*) "Edited Nearest Neighbours". Estratégias de *grid-search* foram aplicadas para a otimização dos hiperparâmetros de cada modelo, utilizando a validação cruzada com k = 5, sendo a área sob a curva ROC (AUC) a métrica da otimização.

| | Total | Teste Positivo | Teste Negativo |
|---|---|---|---|
| Participantes, n (%) | 63855 (100) | 9101 (14.25) | 54754 (85.75) |
| Feminino, n (%) | 38960 (61.01) | 5434 (59.71) | 33526 (61.23) |
| Idade (anos), mediana [IQR] | 43.0 (35 - 54) | 45.0 (36 - 57) | 43.0 (35 - 54) |
| Profissional de Saúde, n(%) | 30647 (47.99) | 3684 (40.48) | 26963 (49.24) |
| Febre | 18741 (29.35) | 4421 (48.58) | 14320 (26.15) |
| Tosse | 32046 (50.19) | 5731 (62.97) | 26315 (48.06) |
| DorGarganta | 27667 (43.33) | 4036 (44.35) | 23631 (43.16) |
| Coriza | 34551 (54.11) | 5287 (58.09) | 29264 (53.45) |
| Mialgia | 28326 (44.36) | 5455 (59.94) | 22871 (41.77) |
| Enjoo | 9257 (14.5) | 1863 (2.05) | 7394 (13.5) |
| Diarreia | 17002 (26.63) | 2914 (32.02) | 14088 (25.73) |
| PerdaOlfato | 17105 (26.79) | 5174 (56.85) | 11931 (21.79) |
| FaltaAr | 1134 (1.78) | 312 (3.43) | 822 (1.5) |
| SemSintomas | 12208 (19.12) | 1022 (11.23) | 11186 (20.43) |

Figura 1: Descrição estatística das variáveis.

A Figura 2 apresenta uma comparação da curva ROC dos 3 modelos utilizados. Pode-se notar que os modelos RF e XG-Boost apresentam desempenho superior à RL. No entanto, são modelos cujos resultados não são facilmente interpretáveis, como acontece com a RL, cujos parâmetros estão diretamente relacionados com a importância dos atributos e permitem o cálculo dos odds ratios.

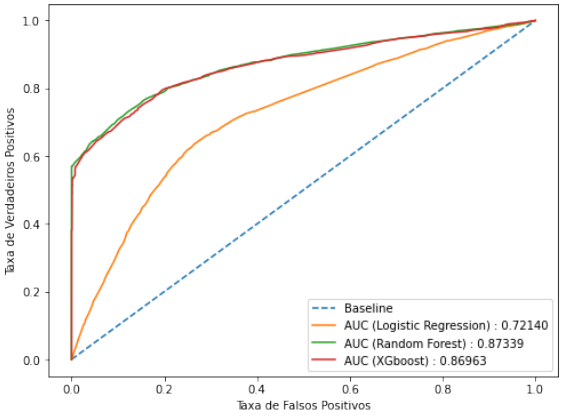


Figura 2: Curva ROC para cada modelo utilizado.

Nesse sentido, utilizamos o método SHAP para calcular a distribuição dos valores Shapley, por atributo, dos métodos RF, XG-Boost e RL, que são representadas na Figura 3, 4 e 5, respectivamente. A Figura 6 apresenta os valores de odds ratios obtidos do modelo de RL com os seus respectivos intervalos de confiança (95%).

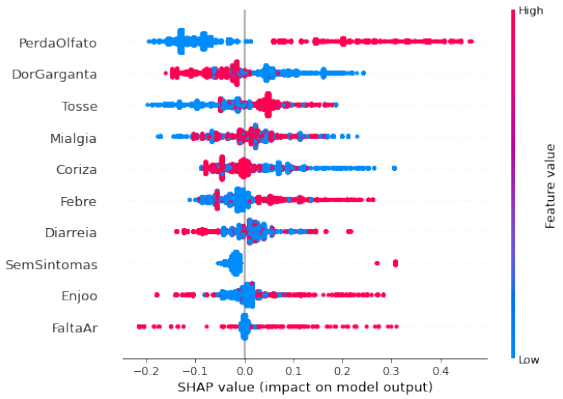


Figura 3: Summary Plot Random Forest.

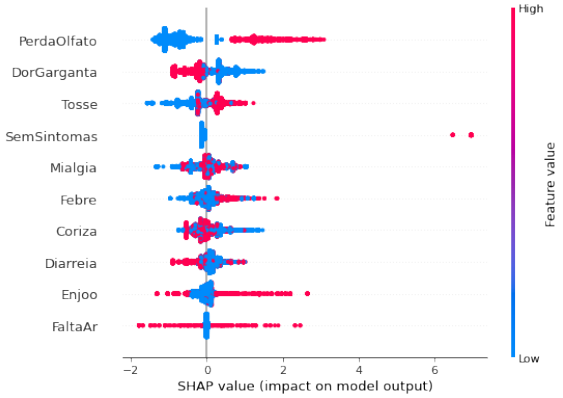


Figura 4: Summary Plot Xgboost.

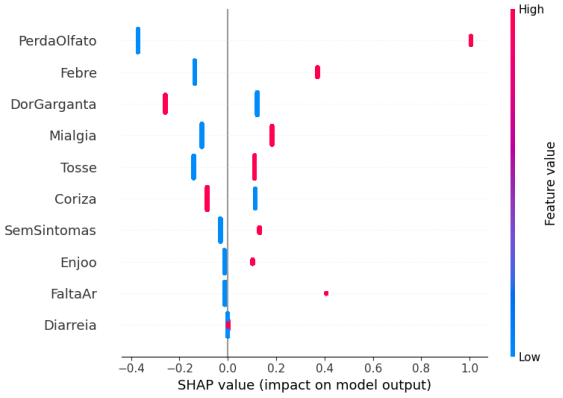


Figura 5: Summary Plot Logistic Regression.

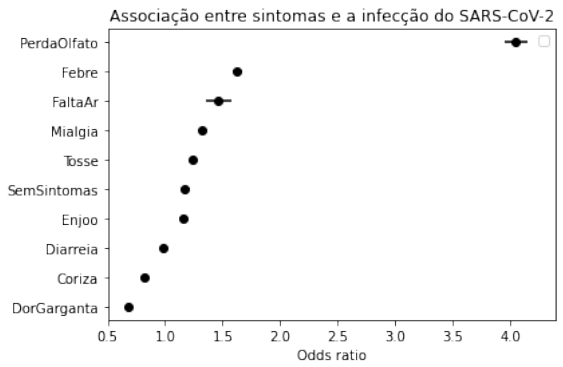


Figura 6: Associação entre sintomas e a infecção do SARS-CoV-2 pelo uso do modelo de Regressão Logística.

Por fim, foi implementado o método LIME para os três modelos de classificação utilizados. Apesar de ser um método local, foi realizada uma agregação global sobre as 10000 primeiras amostras da base de dados para cada modelo, como pode ser observado nas Figuras 7, 8 e 9 .

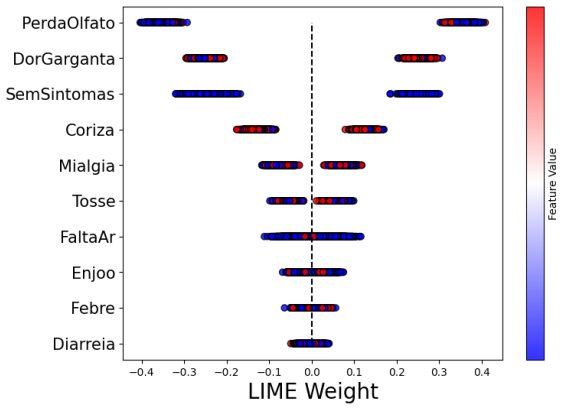


Figura 7: Beeswarm plot Random Forest.

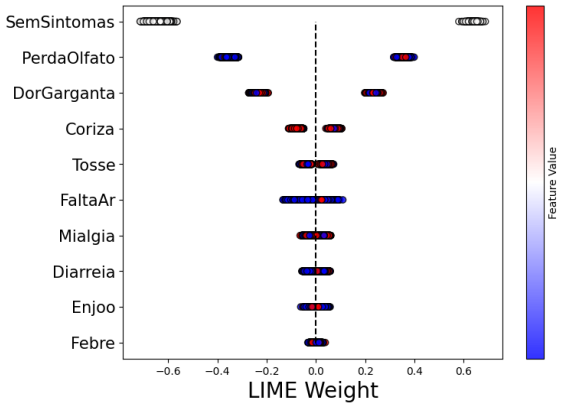


Figura 8: Beeswarm plot Xgboost.

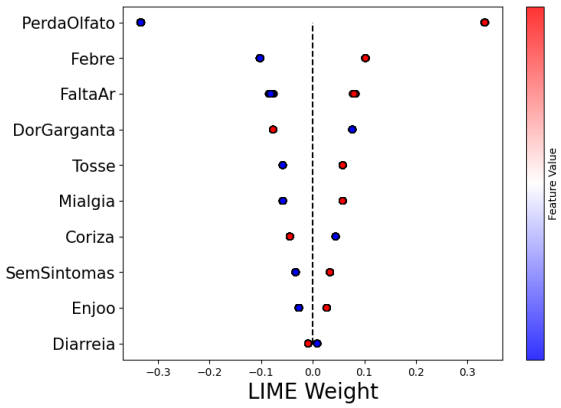


Figura 9: Beeswarm plot Logistic Regression.

Pode-se observar que o sintoma referente à Perda de Olfato (PerdaOlfato) apresentou a maior importância na interpretação em todos os modelos utilizados, com exceção do uso do método Lime no modelo Xgboost, representado na Figura 8. Ademais, a partir dos resultados obtidos para este sintoma, os métodos indicam que a presença da perda de olfato em pacientes aumenta significativamente a probabilidade predita dos mesmos apresentarem o patógeno SARS-COV-2.

Ao analisar o sintoma "DorGarganta", nota-se que os resultados da interpretabilidade também convergem. Apesar deste atributo ser relevante para os métodos SHAP e LIME (apresenta uma alta colocação no ordenamento), a distribuição de cores ao longo da abscissa nas Figuras 3, 4, 5 e 9 indicam que este sintoma reduz a probabilidade dos pacientes apresentarem Covid 19. Tal fato é coerente com o resultado apresentado na Figura 6, pois o odds ratio deste sintoma é aproximadamente 0,68, o que significa que há uma redução de aproximadamente 32% na chance de um paciente testar positivo para Covid 19, quando este sintoma está presente.

# 4 Conclusões

Neste trabalho estudamos dois dos principais métodos interpretáveis utilizados em aprendizado de máquina: SHAP e LIME. Além do estudo teórico, os métodos foram aplicados para um problema de predição de diagnóstico de COVID-19 a partir de sintomas relatados. A partir de uma base de dados pública, os métodos LIME, SHAP e Odds Ratio foram implementados com a linguagem *python* para os modelos RF, RL e Xgboost. Os resultados obtidos permitiram interpretar os modelos *black boxes* e comparar as informações extraídas com modelos interpretáveis à priori, bem como com a literatura médica sobre a classificação de SARS-COV-2 com base em sintomas.

# Referências


[1] Leila F Dantas and et al. App-based symptom tracking to optimize sars-cov-2 testing strategy using machine learning. *PloS one*, 16(3):e0248920, 2021.

[2] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future Hospital Journal*, 6:94–98, 06 2019.

[3] Bert Heinrichs and Simon Eickhoff. Your evidence? machine learning algorithms for medical diagnosis and prediction. *Human Brain Mapping*, 41, 12 2019.

[4] Tyler J Loftus, Patrick J Tighe, Tezcan Ozrazgat-Baslanti, J Parker Davis, Mathew M Ruppert, and Yulong. Ren. Ideal algorithms in healthcare: Explainable, dynamic, precise, autonomous, fair, and reproducible. *PLOS Digital Health*, 1(1):e0000006, 2022.

[5] Christoph Molnar. *Interpretable Machine Learning*. Github, 2 edition, 2022.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[7] Scikit-learn. Machine learning in python. https://scikit-learn.org/stable/. Accessed: 2022-05-17.

[8] Chandan Singh, Keyan Nasseri, Yan Shuo Tan, Tiffany Tang, and Bin Yu. imodels: a python package for fitting interpretable models, 5 2022.