Modelos de Classificação de Aprendizagem de Máquina para Priorização de Teste de COVID-19 no Brasil

Íris Viana dos S. Santana[1], Álvaro Alvares de Carvalho César Sobrinho[1]

Universidade Federal do Agreste de Pernambuco (UFAPE)
Avenida Bom Pastor, s/n.º– Boa Vista – Garanhuns – PE – Brazil

vianasantana21@gmail.com , alvaro.alvares@ufape.edu.br

Abstract. *The aim of this study is to effectively prioritize symptomatic patients for testing COVID-19 in Brazil and thus assist in early detection, addressing problems related to testing and control strategies. 55,676 data were pre-processed, and the chi-square test was performed to confirm the relevance of the following features: gender, health professional, fever, sore throat, dyspnea, cough, coryza, headache, olfactory and taste disorders. Classification models were implemented relying on data sets; supervised learning; and classic algorithms . One of the models with the highest performance was the decision tree (mean accuracy ≥ 89.12%), as it is easy to interpret, it was considered the most adequate.*

Resumo. *O objetivo com este estudo é priorizar efetivamente pacientes sintomáticos para teste de COVID-19 no Brasil e auxiliar na detecção precoce, abordando problemas relacionados à testagem e estratégias de controle. Foram pré-processados 55.676 dados e o teste do qui-quadrado foi usado para confirmar a relevância dos campos: gênero, profissional de saúde, febre, dor de garganta, dispneia, tosse, coriza,dor de cabeça, distúrbios olfatórios e do paladar. Modelos de classificação foram implementados com base nos dados; aprendizagem supervisionada; e algoritmos clássicos. Um dos modelos de maior desempenho foi o de árvore de decisão (média de acurácia ≥ 89,12%), como é de fácil interpretabilidade, foi considerado o mais adequado.*

1. Introdução

A confirmação do primeiro caso de COVID-19 no Brasil ocorreu em março de 2020, e, desde então, existe uma tendência de aumento nos casos confirmados e mortes. Até junho de 2021, o governo brasileiro registrou mais de 16 milhões de casos, com mais de 468.000 mortes. Atualmente, o Brasil é um dos países mais afetados pela COVID-19, com implementação insuficiente de medidas de controle (Monteiro et al. 2020).

Neste estudo, é abordada a priorização de testes de COVID-19 para pacientes que são sintomáticos, auxiliando na detecção precoce da doença. A seguinte Pergunta de Pesquisa (PP) principal foi definida: características demográficas e sintomas que não exigem exames caros podem ajudar efetivamente na priorização de teste para detecção precoce de COVID-19? Dada a PP principal, quatro PPs secundárias foram definidas: (1) quais características demográficas são relevantes para priorização de testes? (2) quais sintomas são adequados para priorização de testes? (3) qual é o modelo de classificação mais adequado para priorização de testes? e (4) quais são os impactos da redução de sintomas relatados na priorização de testes?

Foi realizado o pré-processamento de uma base de dados brutos com informações sobre 55.676 pacientes, com o objetivo de fornecer um modelo de classificação que efetivamente recomenda ou não a priorização de pacientes sintomáticos para o teste de COVID-19 (problema de classificação binária). Algoritmos foram treinados e testados usando bases pré-processadas, compostas por características demográficas e sintomas que não requerem exames caros (Santana et al. 2021). Exames caros foram evitados, dados os níveis de pobreza identificados na população brasileira (Malta et al. 2020).

3. Metodologia

As seguintes atividades foram realizadas: pré-processamento, geração de novas bases de dados, seleção de recursos, validação cruzada com 10 partições, comparações estatísticas e ranqueamento de atributos. Dados brutos de 55.676 brasileiros foram pré-processados para definir novas bases com informações sobre pacientes sintomáticos testados para COVID-19 usando RT-PCR e testes rápidos (anticorpos e antígenos). O teste qui-quadrado foi aplicado nas novas bases para auxiliar na seleção de atributos com $P < 0.01$, verificando a relevância para a tarefa de classificação (Chatterjee et al. 2020).

O método de validação cruzada com 10 partições e cinco repetições foi aplicado para validar os modelos de classificação *Multilayer Perceptron* (MLP), *Gradient Boosting Machine* (GBM), *Decision Tree* (DT), *Random Forest* (RF), *Extreme Gradient Boosting* (XGBoost), *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Logistic Regression with Weak Regularization* (LR) e *Logistic Regression with Strong Regularization* (LRR) usando seis bases. Os resultados médios para as métricas de classificação também foram calculados: *a proporção de classificações que são verdadeiras positivas* (precisão), *a proporção de casos que foram corretamente previstos* (acurácia), *a* proporção *de positivos reais que são corretamente positivos* (*recall)*, *diferença quadrática média entre as probabilidades previstas e os resultados esperados* (pontuação de Brier), *Area Under the Receiver Operating Characteristic* (AUROC) e *Area Under the Precision-Recall Curve* (AUPRC). Os resultados de *recall* foram analisados usando os testes estatísticos de Friedman e Nemenyi durante as comparações dos modelos. O ranqueamento de atributos foi realizado para os modelos com melhor desempenho usando a importância de permutação de atributos.

Os dados brutos utilizados neste estudo foram coletados e disponibilizados pelo órgão público de saúde da cidade de Campina Grande. Tal órgão é informado sobre todos os exames de COVID-19 realizados na cidade. Os funcionários da agência de saúde retiraram a identificação do paciente e os dados disponibilizados foram utilizados para viabilizar este estudo. A base bruta possui atributos categóricos, como profissional de saúde e segurança, etnia, tipo de teste, febre, dor de garganta, dispneia, distúrbios olfativos, tosse, coriza, distúrbios do paladar, dor de cabeça, sintomas adicionais, resultado do teste, comorbidades, situação do teste e descrição de sintomas.

O pré-processamento dos dados e implementação foram realizados usando Python 3.9. A base bruta foi pré-processada aplicando algoritmos de correspondência de *string* para corrigir inconsistências. Um exemplo de inconsistência foi a ocorrência de colunas vazias de sintomas; no entanto, os mesmos sintomas estavam em uma coluna para a descrição geral dos sintomas. Além disso, as seguintes instâncias da amostra total de 55.676 foram removidas devido aos critérios de exclusão: pacientes com testes incompletos ou classificações finais indefinidas (n = 12.929, 23,22%), instâncias duplicadas (n = 251, 0,45%), problemas relacionados à entrada erros (n = 10.408, 18,69%), tipos de teste que não são RT-PCR ou rápidos (n = 771, 1,38%), género indefinido (n = 27, 0,05%) e pacientes que eram assintomáticos (n = 11.269, 20,24%). Os pacientes que eram assintomáticos foram removidos porque as entradas para os algoritmos dependem de características demográficas e sintomas.

Seis bases pré-processadas foram utilizadas: *RT-PCR balanceada*, *RT-PCR desbalanceada*, *Rápido balanceada*, *Rápido desbalanceada*, *Ambos balanceada* e *Ambos desbalanceada*. O particionamento em seis bases, se deu, para analisar com mais detalhes situações em casos reais, por exemplo, dados desbalanceados também foram considerados, sem subamostragem, para melhorar a representatividade dos experimentos e para alcançar um cenário mais próximo de um cenário do mundo real, com mais casos de COVID-19 negativos do que positivos. A separação por tipo de teste foi realizada porque o teste RT-PCR é considerado mais confiável que testes rápidos. Dessa forma, foi analisado separadamente.

4. Resultados

Com base nos valores médios de precisão, acurácia, *recall* e pontuação de Brier obtidos usando a validação cruzada com 10 partições e cinco repetições, foi possível identificar que o modelo de árvore de decisão apresentou desempenho relevante (Tabela 1). Isto pode ser observado para *RT-PCR desbalanceado/balanceado* e *Ambos desbalanceados/balanceados*. Os outros modelos, omitidos por limitação de espaço, apresentaram menores desempenhos quando comparados com o modelo de árvore de decisão.

Tabela 1. Resultados de validação cruzada com 10 partições, para o modelo de árvore de decisão, usando as seis bases de dados.

Bases e Modelos	**Precisão**	**Acurácia**	***Recall***	**Brier**
RT-PCR				
DT, desbalanceada (balanceada)	97.49 (96.50)	96.33 (95.91)	97.04 (95.32)	0.04 (0.04)
Rápido				
DT, desbalanceada (balanceada)	99.37 (95.51)	98.69 (94.59)	99.27 (93.67)	0.01 (0.05)

Ambos				
DT, desbalanceada (balanceada)	95.43 (93.75)	94.79 (89.12)	99.10 (83.87)	0.05 (0.11)

Ao remover atributos de acordo com os resultados do teste qui-quadrado, houve uma diminuição considerável no desempenho dos modelos. Considerar esse cenário é relevante para analisar como os algoritmos se comportam quando os modelos são implementados com um número reduzido de sintomas relatados. Além disso, ao calcular o AUROC usando o RT-PCR, rápido e ambos os cenários de teste, os *trade-offs* entre sensibilidade (taxa de verdadeiro positivo) e probabilidade (taxa de falso positivo) foram identificados, evidenciando as habilidades diagnósticas quando o limiar de discriminação é variado (Figura 1). Os modelos apresentaram alto nível discriminatório para todos os cenários, com as curvas mais próximas do canto superior esquerdo de cada representação gráfica. Porém, para tais cenários, ocorreu uma redução para os modelos KNN e SVM.

Resultados de AUPRC foram usados para verificar os modelos ao lidar com a classe minoritária, analisando o *trade-off* entre precisão e *recall* para diferentes limites de decisão. Nas bases, o rótulo 0 representa o diagnóstico confirmado. DT e XGBoost alcançaram o melhor valor de precisão média (65%) usando a base RT-PCR desbalanceada. Para os demais cenários, os modelos apresentaram valores entre 80% e 96%. Além disso, resultados dos testes de Friedman e Nemenyi foram usados para melhorar a confiança na comparação, com foco nos resultados de *recall,* dados os impactos indesejáveis de falsos negativos. Os resultados indicam que a diferença entre os valores de *recall* é provavelmente real. Para a maioria das bases, a diferença crítica entre LRR/LR (estatisticamente indistinguível) e os outros modelos foi destacada. Dependendo da base, MLP e GBM também foram estatisticamente indistinguíveis, como foi o caso de DT, RF, XGBoost, KNN e SVM.

Os principais atributos dos modelos com os melhores desempenhos (MLP, GBM, DT, RF, XGBoost e SVM) foram identificados usando o método de importância de permutação de atributos. Febre possui valores médios de importância mais altos para quase todos os cenários do que outros sintomas. Gênero e profissionais de saúde estão relacionados a características demográficas relevantes para apoiar a priorização do teste COVID-19 (PP 1). Febre, dor de garganta, dispneia, distúrbios olfatórios, tosse, coriza, distúrbios do paladar e dor de cabeça são sintomas relevantes (PP 2). Para aprimorar os experimentos, os modelos foram combinados usando a estratégia de votação por maioria. Duas combinações foram consideradas para cada base: modelos baseados em árvores (GBM, DT, RF e XGBoost) e outros modelos (MLP, SVM, KNN, LRR e LR). Os resultados médios das métricas usando validação cruzada com 10 partições foram semelhantes aos apresentados acima. Mais detalhes sobre todos os resultados obtidos neste estudo podem ser acessados em Santana et al. (2021).

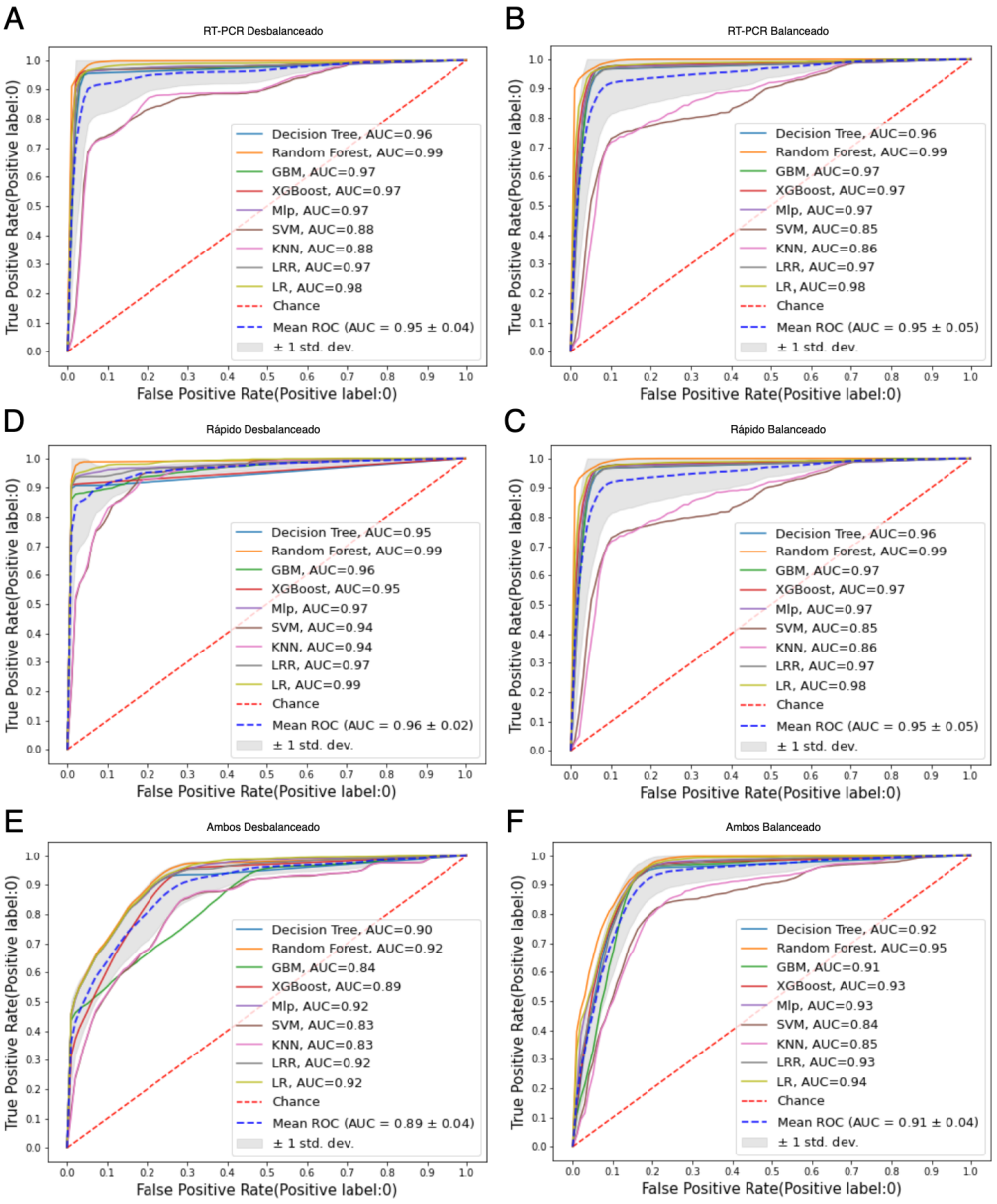

Figura 1. Curvas ROC dos modelos de classificação implementados.

5. Conclusão

Os resultados obtidos são relevantes para demonstrar a viabilidade do uso de modelos para priorização de teste de COVID-19 no Brasil, principalmente com base em sintomas que não requerem exames caros. Com o resultado do teste de qui-quadrado, gênero, profissional de saúde, febre, dor de garganta, dispnéia, tosse, coriza, dor de cabeça, distúrbios olfatórios e do paladar foram consideradas relevantes para priorização de testes de Covid-19.

Ao comparar os modelos usando dados brutos de 55.676 brasileiros, o método de validação cruzada com 10 partições, métricas de classificação e os testes de Friedman e Nemenyi, MLP, GBM, DT, RF, XGBoost e SVM apresentaram os maiores desempenhos com resultados similares. Aplicando a fácil interpretabilidade como critério adicional de comparação, DT foi

considerado o modelo mais adequado (PP 3). Ao remover atributos de acordo com os resultados do teste qui-quadrado, houve uma diminuição considerável no desempenho dos modelos, dessa forma, foi possível perceber que quanto mais sintomas forem informados, o desempenho do modelo será melhor (PP 4).

Os modelos de classificação implementados podem ser a base para os sistemas de *eHealth* e *mHealth* para apoiar profissionais de saúde e formuladores de políticas públicas durante a priorização do teste COVID-19. Como trabalho futuro, para ser aplicado na prática clínica e integrado com o fluxo de trabalho clínico atual, recomenda-se a disponibilidade do modelo de classificação DT e o uso de informações de classificação de recursos por meio de serviços web a serem consumidos por um sistema. Tal sistema deve apresentar resultados de classificação de uma maneira amigável. A fácil interpretação dos modelos é relevante para aumentar a confiança dos profissionais de saúde nos resultados da classificação. Por exemplo, os serviços web podem ser integrados aos sistemas das unidades de saúde públicas brasileiras para priorizar os recursos de teste COVID-19 reduzidos.

Em um cenário real, o número de pacientes assintomáticos com COVID-19 pode ser considerado uma limitação à aplicabilidade dos modelos de classificação. Neste caso, este estudo continua a ser relevante devido aos casos sintomáticos que também necessitam de atenção pelos profissionais de saúde e governo. A avaliação de pacientes sintomáticos também é relevante para prevenir o uso não planejado de recursos de teste COVID-19 devido a outros surtos no Brasil causados por outras doenças virais (por exemplo, dengue). Essas infecções virais apresentam sintomas que podem complicar a decisão dos profissionais de saúde sobre o tipo de teste adequado necessário.

6. Referencias

Chatterjee, A., Gerdes, M. W., Martinez, S. G. (2020). Identification of risk factors associated with obesity and overweight-a machine learning overview. *Sensors (Basel)*, 20(9):2734.

Malta, M., Murray, L., Silva, C. M. F. P., Strathdee, S. A. (2020). Coronavirus in Brazil: the heavy weight of inequality and unsound leadership. *EClinicalMedicine*, 25:100472.

Monteiro, O. M., Fuller, T. L., Brasil, P., Gabaglia, C. R., Nielsen-Saines, K. (2020). Controlling the COVID-19 pandemic in Brazil: a challenge of continental proportions. Nature Medicine, 26(10):1505-1506.

Santana Í. S. S., Silveira, A. C. M., Sobrinho, Á., Silva, L. C., Silva, L. D., Santos, D. F. S., Gurjão, E. C., Perkusich, A. (2021). A Brazilian dataset of symptomatic patients for screening the risk of COVID-19. *Mendeley Data*.

Santana, Í. V. S., Silveira, A. C. M., Sobrinho, Á., Silva, L. C., Silva, L. D., Santos, D. F. S., Gurjão, E. C., Perkusich, A. (2021). Classification Models for COVID-19 Test Prioritization in Brazil: Machine Learning Approach. *Journal of Medical Internet Research*, 23(4).