Aprendizado de Máquina para Predição de Sequelas de Pacientes Com Diagnóstico de COVID-19

Christian Giménez Barañano, Adriano Velasque Werhli

Centro de Ciências Computacionais – Universidade Federal de Rio Grande (FURG)
Av. Itália, km 8 - Campus Carreiros - 96203-900 - Rio Grande – RS – Brasil

christian.baranano@gmail.com, werhli@gmail.com

Abstract. *This study analyzed data from COVID-19 patients treated at the University Hospital of FURG, considering information from the initial hospitalization and the post-recovery follow-up, when sequelae were observed. After data processing and filtering, machine learning techniques were applied to identify the most effective algorithm for predicting sequelae, even with a reduced dataset. The supervised method with cross-validation indicated Naive Bayes as the most suitable. The results aim to support the understanding of the long-term consequences of COVID-19 and the planning of future care.*

Resumo. *Este estudo analisou dados de pacientes com COVID-19 atendidos no Hospital Universitário da FURG, considerando informações da internação inicial e do retorno pós-recuperação, quando apresentaram sequelas. Após tratamento e filtragem dos dados, aplicaram-se técnicas de aprendizado de máquina para identificar o algoritmo mais eficaz na predição de sequelas, mesmo com base de dados reduzida. O método supervisionado com validação cruzada indicou o Naive Bayes como o mais adequado. Os resultados visam apoiar o entendimento das consequências prolongadas da COVID-19 e o planejamento de cuidados futuros.*

1. Introdução

Os vírus acompanham a humanidade há milhões de anos e podem causar doenças graves [Silva, Delatorre, *et al.*, 2023]. Entre eles, os coronavírus destacam-se por sua ampla disseminação entre animais e humanos. O SARS-CoV-2, agente da COVID-19 [Uzunian, 2020], identificado em Wuhan (China) em dezembro de 2019, espalhou-se rapidamente pelo mundo, gerando altas taxas de morbimortalidade [Alves, 2020]. Muitos sobreviventes desenvolveram sintomas persistentes, caracterizando a chamada covid longa, definida pela OMS como a permanência de manifestações clínicas por três meses ou mais após a infecção inicial, sem outra causa aparente. Essa nova condição tem impulsionado pesquisas sobre prevalência, sintomas e tratamento [Segata e Löwy, 2024], sendo mais comuns fadiga, dispneia, distúrbios do sono e dificuldades cognitivas [Alkodaymi, Omrani, *et al.*, 2022]. Diante do grande volume de dados clínicos e da necessidade de prognósticos precisos, técnicas de Inteligência Artificial (IA) despontam

como ferramentas promissoras para identificar padrões e apoiar decisões em saúde. Nesse contexto, este estudo investigou o uso de algoritmos de aprendizado de máquina para prever sequelas pós-COVID em pacientes atendidos no HU-FURG.

2. Metodologia

O estudo utilizou dados de dois subsistemas do HU-FURG: "Fichas COVID" (pacientes internados na fase aguda) e "Pós-COVID" (pacientes com sintomas persistentes no retorno). A base de dados, extraída em formato **.csv** via programa PHP (desenvolvido pelo autor), foi processada na linguagem R Após processado, o conjunto de dados final teve 127 registros e 203 variáveis. O pré-processamento incluiu categorização, discretização de idade, conversão de datas, e limpeza de dados (remoção de colunas irrelevantes/nulas e instâncias isoladas). Foram estabelecidas três modalidades de classificação: duas classes (Positivo/Negativo), três classes (Positivo /Negativo /Astenia) e quinze classes (diversos tipos de sequelas). A divisão da base seguiu proporção de 74% para treino e 26% para teste. Em um segundo momento, foi aplicado o método de validação cruzada K-fold (k=10). Cinco algoritmos de aprendizado de máquina foram avaliados — Árvore de Decisão, Random Forest, Naive Bayes, XGBoost e Rede Neural — com o objetivo de identificar o modelo mais adequado para a predição de sequelas pós-COVID no contexto do HU-FURG. O processamento foi realizado em um computador Desktop com processador AMD Ryzen 7 5700, 32GB de RAM, sistema operacional Windows. A base original consta no servidor principal do HU-FURG.

3. Resultados e Discussão

Os algoritmos de aprendizado de máquina utilizados foram: a **Árvore de Decisão** baseado no princípio “dividir para conquistar”, o método segmenta os dados conforme suas características [Garcia, 2003], apresentando desempenho instável e sensível à variabilidade dos dados e à complexidade do problema. O **XGBoost**, combina iterativamente modelos fracos em um modelo forte [Chen e Guestrin, 2016], mas, apesar da robustez, falhou em algumas partições, sendo excluído. A **Rede Neural** é um campo da inteligência artificial, que busca implementar modelos matemáticos que se assemelhem às estruturas neurais biológicas [Ferneda, 2006]. Obteve melhor desempenho em classificação binária, mas perdeu eficácia em múltiplas classes, especialmente com 15 classes, onde falhou em alguns casos. O **Random Forest** uma coleção de árvores de decisão, utiliza um subconjunto aleatório de atributos para formar as perguntas e um conjunto aleatório de dados de treinamento [Koehrsen, 2017]. Ele superou a Rede Neural em três classes, mas teve uma queda de desempenho com 15 classes. Por fim, o **Naive Bayes** refere-se à construção de um modelo probabilístico Bayesiano para atribuir a probabilidade de classe posterior a uma instância [Berrar, 2018]. Destacou-se na tarefa com 15 classes por ser estável, consistente e gerar resultados em todos os cenários, mesmo sem atingir as acurácias máximas. Para analisar os resultados, foram utilizadas as seguintes métricas: a **Matriz de Confusão**, que apresenta os acertos e erros do modelo,

detalhados em **Verdadeiro Positivo** (classe positiva corretamente identificada), **Falso Positivo** (classe negativa prevista como positiva), **Falso Negativo** (classe positiva prevista como negativa) e **Verdadeiro Negativo** (classe negativa corretamente prevista) [Rodrigues, 2019]. A **Acurácia (Accuracy)** indica a proporção de classificações corretas em relação ao total [Rodrigues, 2019]. A **Precisão (Precision)** identifica quais classificações previstas como positivas estão corretas [Rodrigues, 2019]. O **Recall** identifica quantas das situações de classe positiva esperadas estão corretas [Rodrigues, 2019]. O **F1-Score** é definido como a média harmônica entre precisão e recall [Rodrigues, 2019]. Por fim, foram usadas as médias **Macro** (média aritmética, onde todas as classes contribuem igualmente) [Khalusova, 2022], **Micro** (todas as amostras contribuem igualmente) [Khalusova, 2022] e **Macro Weighted** (média ponderada, onde a contribuição de cada classe é ponderada pelo seu tamanho) [Khalusova, 2022].

A Figura 1 apresenta as médias das principais métricas de desempenho obtidas pelos algoritmos Árvore de Decisão (AD), Naive Bayes (NB), Random Forest (RF) e Rede Neural (RN). Observa-se que, embora a Rede Neural (RN) tenha alcançado o maior valor pontual em *Precision_Weight*, esse resultado apresentou alta variabilidade, indicando baixa estabilidade. O Random Forest (RF) exibiu desempenho consistente, porém inferior em algumas métricas de recall. De forma geral, considerando o conjunto das médias, o Naive Bayes (NB) apresentou o desempenho mais equilibrado e adequado entre as métricas avaliadas, sugerindo melhor capacidade de generalização para os dados analisados. O uso do método de validação cruzada K-fold confirmou a estabilidade do modelo, reforçando sua adequação para cenários com bases reduzidas e alta variabilidade. Esses resultados sugerem que o Naive Bayes é uma abordagem promissora para apoiar o monitoramento clínico e o planejamento de cuidados pós-COVID no contexto hospitalar.

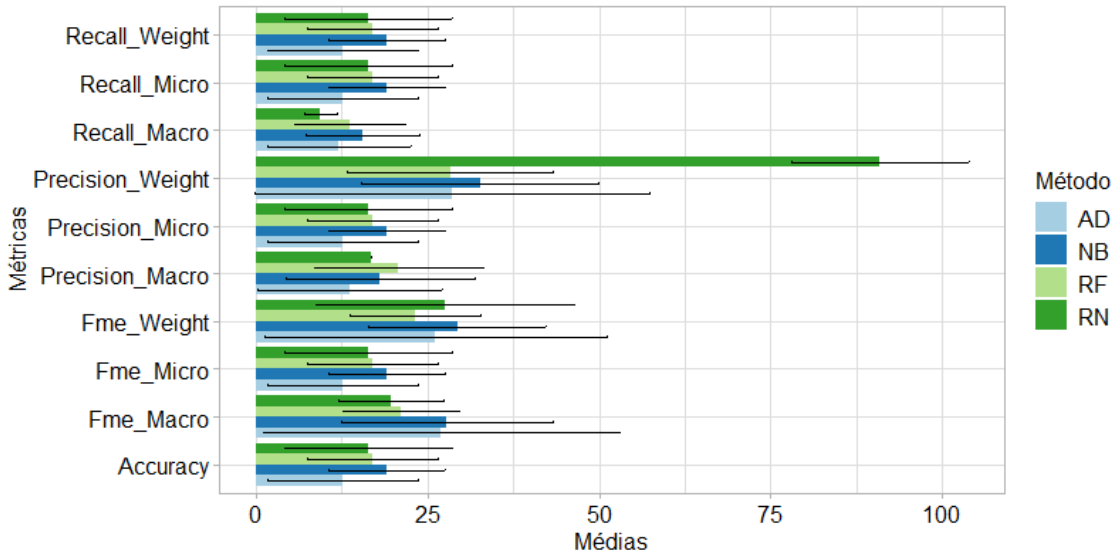

Figura 1. Comparativo das Partições – 15 classes. As barras pretas indicam o desvio padrão das médias. Legenda: AD – Árvore de Decisão, NB – Naive Bayes, RF – Random Forest e RN – Rede Neural

Referências

Alkodaymi, M. S., Omrani, O. A., Fawzy, N. A; Shaar, B. A., Almamlouk, R., Riaz, M., Obeidat, M., Obeidat, Y., Gerberi, D., Taha, R. M., Kashour, Z., Kashour, T., Berbari, E. F., Alkattan, K. and Tleyjeh, I. M. (2022) ” Prevalence of post-acute COVID-19 syndrome symptoms at different follow-up periods: a systematic review and meta-analysis”. https://pubmed.ncbi.nlm.nih.gov/35124265/

Alves, J. E. D..(2020) ”O avanço da pandemia de Covid-19 no mundo e no Brasil no mês de março”. https://www.ihu.unisinos.br/categorias/597706-o-avanco-da-pandemia-de-covid-19-no-mundo-e-no-brasil-no-mes-de-marco

Berrar, D. B. (2018) “Theorem and Naive Bayes Classifier”. In: RANGANATHAN, S.; NAKAI, K.; SCHONBACH, C. Encyclopedia of Bioinformatics and Computational Biology. [S.l.]: Elsevier, v. 1, 2018. p. 403-412

Chen, T. Q. and Guestrin, C. (2016) “XGBoost: A Scalable Tree Boosting System”. KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining , San Francisco, p. 753 - 794, jun. 2016.

Ferneda, E. (2006) “Redes neurais e sua aplicação em sistemas de recuperação de informação”. Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto , São Paulo, Brasil. https://www.scielo.br/j/ci/a/SQ9myjZWLxnyXfstXMgCdcH/#

Garcia, S. C. (2003) “O Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde”. Universidade Federal do Rio Grande do Sul. Porto Alegre, p. 14. https://lume.ufrgs.br/handle/10183/4703

Khalusova, M. (2022) “Machine Learning Model Evaluation Metrics part 2**:** Multi-class Classification” https://www.mariakhalusova.com/posts/2019-04-17-ml-model-evaluation-metrics-p2/

Koehrsen, W. (2017) “Random Forest Simple Explanation”. Medium. https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

Rodrigues, V. (2019) “Métricas de Avaliação: acurácia, precisão, recall… quais as diferenças?” Medium, https://vitorborbarodrigues.medium.com/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-recall-quais-as-diferen%C3%A7as-c8f05e0a513c

Segata, J. e Löwy, I. (2024) “Covid longa, a pandemia que não terminou”, Horizontes Antropológicos https://doi.org/10.1590/1806-9983e700601

Silva, A.S., Delatorre, E. O., Leon, L. A. A., Azevedo, S. S. D., Leite, T. C. N. F. e Paula, V. S. (2023) “Tópicos em Virologia”, Editora FIOCRUZ.

Uzunian, A. (2020) “Coronavirus SARS-CoV-2 and Covid-19”. https://www.scielo.br/j/jbpml/a/Hj6QN7mmmKC4Q9SNNt7xRhf/?lang=pt#