Previsão de Casos de COVID-19 no Brasil com Aprendizagem de Máquina Automatizada

Leticia Magalhães Mendes & Elloá B. Guedes

***Resumo*— O presente trabalho considera a utilização de Aprendizagem de Máquina Automatizada para lidar com a previsão antecipada de novos casos de COVID-19 no Brasil. Utilizando uma abordagem de programação genética e validação cruzada do tipo *Time Series Split*, foi possível obter um regressor por Vetores de Suporte Linear com R^2 médio igual a $0{,}4257$ e máximo igual a $0{,}975$, isto é, que melhora suas previsões conforme mais dados tornam-se disponíveis.**

***Palavras-Chave*— Aprendizagem de Máquina; Séries Temporais; COVID-19.**

***Abstract*— The present work aims at using Automated Machine Learning to forecast new COVID-19 cases in Brazil. By using a Genetic Programming approach combined with Time Series Split cross-validation, it was possible to identify a Linear Vector Support regressor with average R^2 of 0.4257 and maximum equal to 0.975, i.e., that improves its performance as more data becomes available.**

***Keywords*— Machine Learning; Time Series; COVID-19.**

I. Introdução

A pandemia do COVID-19 (*Coronavirus disease*) emergiu como um fardo avassalador para a Saúde, tendo acometido, até 2 de Maio de 2021, um total de $151.812.556$ indivíduos em âmbito mundial, causando $3.186.817$ mortes [1], das quais 414.399 ocorreram no Brasil [2]. O combate à pandemia da COVID-19 é um desafio contínuo de proporções mundiais. Emergências de saúde pública como essa exigem uma resposta eficaz em tempo preciso. Neste contexto, até mesmo esforços da Ciência de Dados têm sido evocados, tais como para elaboração de soluções de Vigilância Epidemiológica, de Imagiologia Médica, dentre diversas outras [3].

O uso de modelos de previsão da evolução da pandemia do COVID-19 é essencial para auxiliar a tomada de decisão governamental eficaz, para gerenciar os recursos da cadeia de suprimentos e para implementar ou manter políticas de distanciamento social [4]. Atualmente, as técnicas de Aprendizado de Máquina (ML, do inglês *Machine Learning*) são usadas em todo o mundo para tais previsões em virtude de sua acurácia. No entanto, treinar bons modelos de ML e efetuar o ajuste de seus parâmetros e hiperparâmetros são tarefas desafiadoras quando existem poucos dados disponíveis, o que é um cenário realístico diante do surto de uma nova doença [5]. Ademais, ressalta-se que a utilização de um único modelo para todos os cenários é pouco provável, especialmente diante da qualificação de dados, da variabilidade populacional, da desigualdade social e das ações de gestão públicas em cada localidade [6].

Grupo de Pesquisas em Sistemas Inteligentes, Laboratório de Sistemas Inteligentes, Escola Superior de Tecnologia, Universidade do Estado do Amazonas, Av. Darcy Vargas, 1200 – Manaus – Amazonas, {lmm.eng18, ebgcosta}@uea.edu.br . As autoras agradecem o apoio financeiro provido pela FAPEAM por meio do Edital PPP 004/2017 e do Programa PAIC/FAPEAM/UEA 2020-2021. EBG agradece também o apoio financeiro provido pelo projeto Edital 003/2020 – SEECT/FAPESQ/PB.

Ao considerar as recentes contribuições da literatura na previsão da pandemia do COVID-19, conforme reportado em *surveys* recentes [5], [7], percebe-se uma grande variedade na modelagem da tarefa de previsão, nas fontes de dados, nas variáveis endógenas e exógenas consideradas, nos modelos adotados e nas regiões abrangidas. Embora ML figure como uma das estratégias mais utilizadas, a performance de muitos modelos de ML é extremamente sensível a um grande conjunto de decisões de projeto (tipo de modelo, configuração de parâmetros e hiperparâmetros, etc.) e encontrar o melhor modelo para um dado problema ainda permanece uma questão em aberto [8].

A Aprendizagem de Máquina Automatizada (AutoML, do inglês *Automated Machine Learning*) consiste em um conjunto de métodos, técnicas e tecnologias que visam automaticamente determinar uma melhor abordagem para um certo problema de aprendizado, de maneira objetiva, orientada à dados e automatizada [9]. AutoML não se propõe a substituir especialistas em ML, mas sim a viabilizar técnicas que automatizam a seleção de modelos e a otimização de hiperparâmetros, por exemplo, como uma maneira de alcançar um novo patamar de desempenho em tarefas de aprendizado [10].

Nesse contexto, o objetivo deste trabalho é apresentar os resultados de uma abordagem baseada em AutoML aplicada ao problema da previsão de novos casos de COVID-19 no Brasil com uma semana de antecedência. Para validar e avaliar a solução proposta, considerou-se dados nacionais a respeito da pandemia e uma estratégia de validação cruzada com séries temporais. Para apresentar o que se propõe, o presente artigo está organizado como segue: a Seção II apresenta a metodologia proposta; a Seção III contempla os resultados obtidos e sua discussão; e, por fim, a Seção IV contempla as considerações finais e perspectivas futuras.

II. Materiais e Métodos

Os dados experimentais para este problema foram coletados do Painel Coronavírus do Ministério da Saúde, contemplando o período de 25 de fevereiro de 2020 até 21 de abril de 2021 [2]. Considerou-se apenas os casos acumulados diariamente em nível nacional, correspondendo a um total de 422 observações. Os dados foram organizados segundo uma Série Temporal indexada pelos dias desde a primeira observação.

A previsão de casos de COVID-19 foi modelada como uma tarefa de regressão mediante Aprendizado Supervisionado, em que deseja-se antever o total acumulado de casos de COVID-19 com uma semana de antecedência. Para tanto, considerou-se como atributos preditores (variáveis dependentes) o número de casos no dia i bem como os novos casos diagnosticados nos dois dias anteriores. Não foram utilizados mais atributos preditores sob o risco de haver descontinuidades nas observações da série. Assim, tem-se como objetivo aprender uma função f que

mapeia X_i, $\Delta(X_i, X_{i-1})$ e $\Delta(X_{i-1}, X_{i-2})$ em X_{i+7} (variável independente). Organizando os dados disponíveis segundo esta tarefa, tem-se um total de 413 exemplos disponíveis sobre o problema.

Para avaliar a tarefa de regressão considerou-se o coeficiente de determinação R^2. Esta métrica expressa a quantidade da variância dos dados que é explicada pelo modelo, com valor máximo igual a 1 [11]. Esta avaliação será realizada no contexto de uma validação cruzada do tipo *Time Series Split*, a qual considera a dinâmica de uma série temporal em que, a princípio, poucos dados estão disponíveis para prever o futuro, mas à medida que o fenômeno começa a ser observado, novos exemplos vão sendo progressivamente acrescidos ao conjunto de treino e o horizonte de previsão avança em direção ao futuro [12]. Neste caso, o número de splits foi $k = 58$, correspondente ao número de semanas epidemiológicas existentes nos dados experimentais.

Uma vez estabelecidos os dados experimentais, a métrica de desempenho e a estratégia de validação das soluções candidatas, partiu-se para a obtenção do modelo para a regressão com o uso de AutoML. Para este fim, considerou-se a biblioteca TPOT, a qual é uma ferramenta baseada em árvores para otimização de *pipelines* [13]. Ela baseia-se em uma abordagem de programação genética para propor uma população de regressores (Árvores de Decisão, Florestas Aleatórias, Regressão Logística, dentre outros) com seus respectivos parâmetros e hiperparâmetros, a qual vai evoluindo com operações de seleção, cruzamento e mutação, o que culmina em um modelo particular e sua configuração que promovem um bom desempenho na tarefa original [14]. Além da seleção do modelo, o *pipeline* resultante do TPOT pode conter operadores de seleção e pré-processamento de atributos [10]. No escopo deste trabalho, considerou-se uma população inicial de 150 regressores distintos e 15 épocas de evolução.

III. Resultados e Discussão

Após a implementação da metologia proposta com a linguagem de programação Python em um servidor computacional com processador Intel Core i7 3,7 GHz, 32 GB de memória principal, 960 GB de memória secundária de estado sólido e 2 placas gráficas NVIDIA GTX 1080 Ti com 11 GB cada, o *pipeline* resultante considera o regressor por Vetores de Suporte Linear com $C = 0{,}01$, $\varepsilon = 0{,}0001$, tol $= 0{,}001$ e função de perda tipo L_2 como o mais adequado à tarefa de previsão proposta.

No cenário experimental, a média do R^2 nos *splits* foi de $0{,}4257 \pm 0{,}9591$ e a mediana igual a $0{,}7077$, com mínimo igual a $-5{,}6005$ e máximo igual a $0{,}975$. A observação de um valor de desempenho negativo significa que o modelo foi arbitrariamente ruim. Estas observações, porém, aconteceram no começo do fenômeno, em que os dados de treinamento são escassos. Em contrapartida, no melhor *split*, foi observado um desempenho muito próximo de um modelo perfeito, ou seja, demonstra a capacidade da solução proposta de melhorar satisfatoriamente seus resultados conforme a evolução da pandemia. A título de ilustração das qualidades preditivas do regressor identificado, a Fig. 1 mostra 80% dos dados utilizados para treinamento e as previsões nos 20% dos dados mais recentes.

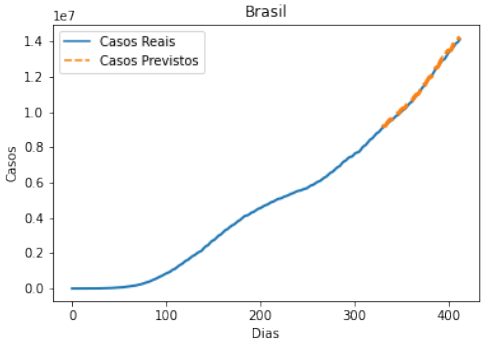

Fig. 1. Previsões de novos casos da COVID-19 no Brasil com o modelo de Regressão por Vetores de Suporte Linear.

IV. Considerações Finais

Os resultados obtidos neste trabalho mostram o potencial de uso de técnicas de AutoML na análise preditiva de novos casos de COVID-19 no Brasil, colaborando para mitigar os impactos da pandemia na sociedade. Em trabalhos futuros, almeja-se analisar comparativamente os resultados obtidos com a literatura, ampliar o espaço da busca na aprendizagem automatizada e efetuar uma análise mais detalhada da convergência dos modelos.

Referências

[1] OMS, "COVID-19 weekly epidemiological update," Organização Mundial de Saúde, Tech. Rep. 4 de Maio de 2021, 2021.

[2] Brasil, "Ministério da saúde – guia de vigilância epidemiológica do COVID-19," 2020, disponível em at https://covid.saude.gov.br/. Acessado em 6 de agosto de 2021.

[3] E. B. Guedes, C. M. S. Figueiredo, and T. E. de Melo, "Esforços da ciência de dados contra a COVID-19 – recursos, chamados à ação e desafios," Secretaria de Desenvolvimento Econômico, Ciência, Tecnologia e Inovação do Governo do Estado do Amazonas & Sociedade Brasileira para o Progresso da Ciência, Manaus, Amazonas, Tech. Rep. Nota Técnica COVID-19 No. 002 de 28/04/2020, 2020.

[4] K. Nikolopoulos, S. Punia, A. Schäfers, C. Tsinopoulos, and C. Vasilakis, "Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions," *European Journal of Operational Research*, vol. 290, no. 1, pp. 99–115, Apr. 2021. [Online]. Available: https://doi.org/10.1016/j.ejor.2020.08.001

[5] G. R. Shinde, A. B. Kalamkar, P. N. Mahalle, N. Dey, J. Chaki, and A. E. Hassanien, "Forecasting models for coronavirus disease (COVID-19): A survey of the state-of-the-art," *SN Computer Science*, vol. 1, no. 4, Jun. 2020.

[6] E. A. Junior, "Covid-19: Desafios para modelagem e análise de dados," *J. Health Inform.*, vol. 12, no. 2, pp. 1–2, 2020.

[7] I. Rahimi, F. Chen, and A. H. Gandomi, "A review on COVID-19 forecasting models," *Neural Computing and Applications*, Feb. 2021. [Online]. Available: https://doi.org/10.1007/s00521-020-05626-8

[8] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. Estados Unidos: Pearson Prentice-Hall, 2009.

[9] F. Hutter, *Automated machine learning: methods, systems, challenges*. Cham, Switzerland: Springer, 2019.

[10] D. Radecic, *Machine Learning Automation withTPOT: build, validate, and deploy fully automated machine learning models with Python*. Packt Publishing Limited, 2021.

[11] A. C. Cameron and F. A. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models," *Journal of Econometrics*, vol. 77, no. 2, pp. 329–342, 1997.

[12] C. N. Bergmeir, "New approaches in time series forecasting: methods, software and evaluation procedures," Ph.D. dissertation, Universidad de Granada, Espanha, 2013.

[13] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.

[14] R. S. Olson and J. H. Moore, "TPOT: A tree-based pipeline optimization tool for automating machine learning," in *Automated Machine Learning*. Springer International Publishing, 2019, pp. 151–160. [Online]. Available: https://doi.org/10.1007/978-3-030-05318-5_8