INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO NO NÚMERO DE CASOS DE COVID-19: APERFIÇOANDO O MODELO PREDITIVO COM NOVAS VARIÁVEIS

Glauber Luiz de Moura Moraes[1]; Júnia Ortiz[2]; Erick Giovani Sperandio Nascimento[3]
[1] Bolsista; HPC FAPESB; glauber.ms@fbter.org.br
[2] Bolsista; Centro de Competência em Inteligência Artificial – SENAI CIMATEC; junia.matos@fbter.org.br
[3] Centro Universitário SENAI CIMATEC; Salvador-BA; erick.sperandio@fieb.org.br

RESUMO

Este trabalho tem como objetivo descrever um modelo de *deep learning* desenvolvido para predição da quantidade de casos de COVID-19, avaliando sua capacidade preditiva em duas configurações distintas no que diz respeito às variáveis de entrada: uma utilizando apenas os dados diários de casos, mortes e medidas de mitigação; e outra acrescentando uma série de variáveis socioeconômicas à entrada do modelo (número de leitos, quantidade de pessoas vacinadas, população, expectativa de vida, dentre outras). Os dados são extraídos dos repositórios online disponibilizados pela Universidade John Hopkins e Universidade de Oxford. O desempenho do modelo com e sem as novas variáveis foi avaliado quantitativamente com a utilização das métricas MSE, RMSE, MAE, Pearson R, Fator de 2 e NMSE. A rede treinada com a inclusão de variáveis socioeconômicas apresentou desempenho superior ao modelo treinado sem o acréscimo destas variáveis, demonstrando que esta inclusão confere ganho no aprendizado do modelo.

PALAVRAS-CHAVE: Inteligência Artificial, COVID-19, predição.

1. INTRODUÇÃO

Os estudos realizados na área de inteligência artificial (IA) podem minimizar o risco de propagação de doenças e indicar estratégias a serem adotadas para a segurança e a saúde da população[1]. Neste sentido, este estudo visa examinar um conjunto de dados socioeconômicos, buscando avaliar a inserção de novas variáveis no aumento da capacidade preditiva de um modelo de inteligência artificial desenvolvido para prever o comportamento da pandemia de COVID-19 para trinta dias à frente. A principal contribuição do experimento aqui apresentado é o incremento no desenvolvimento de modelos de predição de séries temporais para quantidade de casos e óbitos decorrentes da COVID-19 com a utilização de variáveis socioeconômicas, visto que os modelos construídos e encontrados na literatura comumente utilizam apenas os próprios dados de interesse (número de casos e mortes) como variáveis de entrada.

2. METODOLOGIA

Para a realização da predição da série temporal com os dados da COVID-19 (casos confirmados e mortes), foi utilizado um modelo *Ensemble* (ao qual chamamos de Metamodelo), a partir da combinação de dois modelos de base construídos com as redes LSTM (long short-term memory)[4] e uma rede híbrida CNN+LSTM, com camadas convolucionais e recorrentes. O modelo *Ensemble* foi avaliado quanto à sua capacidade preditiva em duas configurações distintas no que diz respeito às suas variáveis de entrada: uma utilizando apenas os dados diários de casos, mortes e medidas de mitigação; e outra acrescentando uma série de variáveis socioeconômicas à entrada do modelo (número de leitos, quantidade de pessoas vacinadas, população, expectativa de vida, dentre outras). Os dados utilizados são públicos e foram coletados no site da Universidade de Oxford[2]. Essa base de dados apresenta atualmente 59 variáveis, dentre as quais, estão incluídas: número de casos confirmados, mortes, testes, vacinados, população do país/região, densidade populacional e expectativa de vida.

As seguintes variáveis de entrada foram adicionadas ao modelo: “new_cases” (Novos casos confirmados), “total_cases” (Total de casos confirmados), “new-deaths” (Novas mortes atribuídas a COVID-19), “total_deaths” (Total mortes atribuídas a COVID-19), “total_cases_per_milion” (Total de casos por um milhão de pessoas), “total_deaths_per_milion” (total de mortes atribuídas a COVID-19 por milhão de pessoas), “new_deathes_per_million” (Novas mortes atribuídas a COVID-19 por um milhão de pessoas), “new-tests” (Novos testes para COVID-19 por mil pessoas), “total_tests” (Total de testes para COVID-19), “total_tests_per_thousand” (Total de testes para COVID-19 por mil pessoas), “new-tests_per_thousand” (Novos testes para COVID-19 por mil pessoas), total_tests_per_thousand” (Total testes para COVID-19 por mil pessoas), “new_tests_per_thousand” (Novos testes para COVID-19 por mil pessoas), “total_vaccinations” (Número total de doses administradas, contendo uma única dose), “people_vaccinations” (Quantidade de pessoas vacinadas), “people_fully_vaccionactions” (Quantitativo de pessoas que receberam todas as doses), “new_vaccinations” (Novas doses administradas), “new_vaccinations_smoothed” (Novas doses

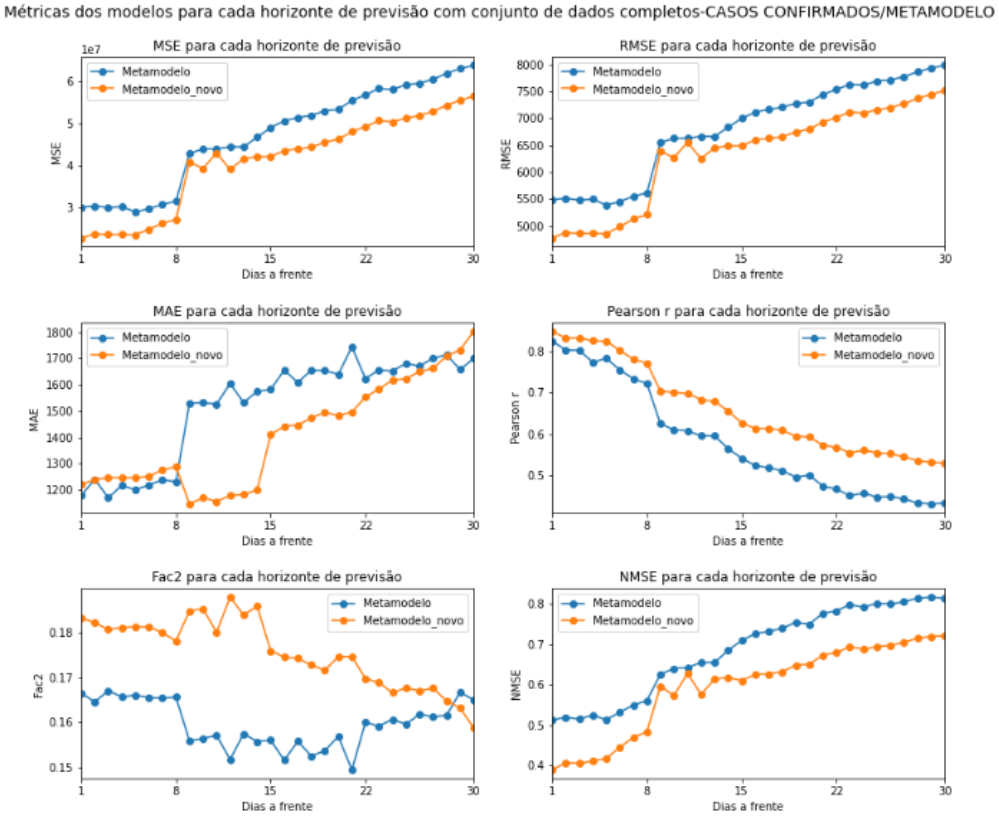
administradas suavizadas), “total_vaccinations_per_hundred” (Proporção de doses administradas por cento da população), “people_ vaccinations_per_hundred” (Proporção de pessoas vacinadas por cento da população), “people_fully_vaccinations_per_hundred” (Proporção de pessoas totalmente vacinadas em cada 100 da população total), “new_fully_vaccinations_smoothedper_per_milion” (Proporção de novas doses administradas por milhão de pessoas), “population” (População em 2020), “population_density” (densidade populacional no ano mais recente), “median_age” (Idade média da população com projeção da ONU em 2020), “age_65_older” (Parcela da população com 65 anos ou mais no ano mais recente), “gdp_per_capita” (Produto interno bruto em paridade com poder de compra disponível no ano mais recente), “life_expectancy” (Expectativa de vida ao nascer em 2019) , “human_development_index” (Medida resumida do índice de desenvolvimento humano).

A capacidade preditiva do modelo com e sem as variáveis socioeconômicas foi avaliada com a utilização das seguintes métricas: MSE (Mean Squared Error), RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Pearson R (Correlação de Pearson), Fac 2 (Fator de 2) e NMSE (Normalized Mean Squared Error).

3. RESULTADOS E DISCUSSÃO

A figura 1 apresenta os resultados das métricas de avaliação dos erros obtidos das predições realizadas pelo modelo sem as variáveis socioeconômicas (Metamodelo) e do modelo com as variáveis socioeconômicas (Metamodelo_novo). Importante destacar que, considerando as métricas de erro (MSE, RMSE, MAE e NMSE), os melhores desempenhos são apresentados pelos modelos com valores menores. Para as métricas Pearson r e Fac2, os melhores modelos são aqueles que apresentam valores mais próximos de 1.

Figura1. Métricas do modelo de previsão do conjunto de dados completo



Conforme pode ser verificado a partir dos gráficos, para as métricas MSE, RMSE, MAE e NMSE o metamodelo novo apresentou um erro menor comparado com o metamodelo sem as novas variáveis. Para

as métricas Fator de 2 e Pearson R, os valores apresentados pelo metamodelo com as variáveis socioeconômicas foram maiores que os valores apresentados pelo metamodelo antigo (o que, neste caso, também indica que o modelo com as novas variáveis teve melhor performance). Verificamos, portanto, que o metamodelo novo apresentou desempenho superior ao metamodelo antigo, conforme indicado por todas as métricas utilizadas para a avaliação.

O resultado do comparativo realizado demonstra o benefício da utilização das novas variáveis de entrada para a predição de casos e óbitos decorrentes de COVID-19, o que leva à importância da consideração de variáveis distintas às comumente utilizadas nas predições de séries temporais neste contexto. Ainda que as variáveis socioeconômicas utilizadas não sejam as principais variáveis envolvidas no fenômeno da pandemia (como as quantidades de casos e óbitos), elas podem explicar, por exemplo, a diferença do comportamento da curva apresentada por diferentes regiões, ainda que as medidas de contingenciamento adotadas sejam similares.

4. CONSIDERAÇÕES FINAIS

O presente trabalho apresentou um modelo preditivo para casos e óbitos decorrentes de COVID-19 treinado com a inclusão de uma série de variáveis socioeconômicas (além das variáveis quantidade de casos, óbitos e medidas de mitigação) com desempenho superior ao mesmo modelo treinado sem as variáveis socioeconômicas. O ganho conferido por esta inclusão no aprendizado do modelo aponta a relevância da utilização de variáveis socioeconômicas no desempenho de um modelo de Inteligência Artificial para previsão do comportamento da pandemia de COVID-19. Os resultados contribuem para o desenvolvimento de melhores ferramentas para a compreensão e previsão de crises como a que estamos vivendo, ajudando a sociedade em geral e as entidades públicas na construção de medidas de enfrentamento de pandemias como esta.

Agradecimentos

A todos os colaboradores do Centro de Referência em Inteligência Artificial – SENAI/CIMATEC e aos colaboradores do Centro de Supercomputação para Inovação Industrial – SENAI/CIMATEC. À FAPESB (fundação de Amparo à Pesquisa do estado da Bahia) pelo auxílio financeiro que possibilitou dedicação ao trabalho. Ao professor Dr Erick Giovani Sperandio Nascimento pela disponibilidade e orientação, possibilitando aprimoramento de conhecimento na área.

5. REFERÊNCIAS

[1] CHIMMULA, V. K. A. **Artificial Intelligence in COVID-19 Related Events Prediction: A Brief Review**. 19 dezembro de 2020. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/21889/17468, acessado em: 01 de abril de 2021 às 20:30.
[2] University of Oxford, **Data on COVID-19**. 08 outubro de 2020. Disponível em: https://github.com/owid/covid-19-data/tree/master/public/data.
[3] CHIMMULA, V. K. A. **Artificial Intelligence in COVID-19 Related Events Prediction: A Brief Review**. 19 dezembro de 2020. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/21889/17468, acessado em: 01 de abril de 2021 às 20:30.
[4] SÁ, G. C. B. SILVA, A. V. LIMA, N. C. A. NASCIMENTO, S. M. **Artificial Intelligence in COVID-19 Related Events Prediction: A Brief Review**. 19 dezembro de 2020. Disponível em: https://www.brazilianjournals.com/index.php/BRJD/article/view/21889/17468.