Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Predição de Surtos de Influenza Usando Aprendizado de Máquina: Um Estudo com Dados da América Latina

Daniela P. Paula[1]
UERJ/ENCE, Rio de Janeiro, RJ
Mayumi D. Makimoto[2], Valdir S. Ermida[3], Fernanda G. F. Salvador[4], Bruno R. da Silva[5], Daylene T. S. Barbosa[6]
INI/Fiocruz, Rio de Janeiro, RJ
Leonardo Gargano[7]
UFRJ, Rio de Janeiro, RJ
Arthur O. C. Schilitz[8]
INPI, Rio de Janeiro, RJ
Wagner M. Junior [9]
UFMG, Belo Horizonte, MG
Ian Miles[10]
UM, Manchester, UK

Influenza é uma doença sazonal de ocorrência mundial com impacto significativo na morbi-mortalidade, especialmente em países de baixa e média renda [1, 3]. Desde o seu aparecimento inicial em 1918, o vírus da influenza sofreu uma evolução genética, melhorando a sua capacidade de invadir as células hospedeiras [3]. A primeira pandemia de influenza H1N1 ocorreu em 1918. Em 1957, H1N1 foi substituída pela estirpe H2N2, e em 1968, surgiu a H3N2. Mais recentemente, em 2009, o H1N1 reemergiu como a primeira pandemia do século XXI, e permaneceu prevalecente até 2019, quando o mundo então enfrentou a grande pandemia causada pelo coronavírus SARS-CoV-2 [3], [1].

O objetivo deste estudo é avaliar a intensidade de epidemias sazonais de influenza na América Latina e identificar possíveis fatores relacionados dentre variáveis socioeconômicas, ambientais, e características de acesso a saúde utilizando técnicas de aprendizagem de máquina como a random forest e SVM (Support Vector Machine), e regressão logística [2].

Foram selecionados os países da América Latina: Argentina, Bolívia, Brasil, Chile, Colômbia, Equador, Guiana, Paraguai, Peru, Suriname, Uruguai e Venezuela. Dados relativos ao número de casos de influenza foram coletados no site da Organização Mundial da Saúde (OMS), nas bases FLUNET e FLUID. Foram coletados os números de casos por semana epidemiológica no período de 2011 a 2024. O tamanho da população anual foi coletado do site do Banco Mundial, no período de 2006 a 2024.

[1]danielapopaula@gmail.com
[2]mayumi.wakimoto@ini.fiocruz.br
[3]valdirermida@gmail.com
[4]afernandasalvador@gmail.com
[5]bruno.rosa@ini.fiocruz.br
[6]dayleneticiane@gmail.com
[7]leogargano@gmail.com
[8]artmestre@gmail.com
[9]meira@dcc.ufmg.br
[10]i.d.miles@manchester.ac.uk

O número de casos mensal de influenza foi calculado somando o número de casos em cada semana epidemiológica do mês. A partir daí foi calculada a taxa de ocorrência mensal por 10.000 habitantes. Também foram calculadas as variações mensais nas taxas, em percentual, entre um mês e seu consecutivo, para cada um dos anos observados. Foram criadas variáveis dicotômicas a partir das variações mensais, indicando, para cada país, quando ocorreu ou não em um ano uma variação maior ou igual a 60%, maior ou igual a 70%, maior ou igual a 80% nas taxas de ocorrência por 10.000 habitantes.

Dados coletados do projeto Our World in Data originaram as variáveis preditoras: número de passageiros que deixaram o país via transporte aéreo, anual de 2006 a 2020; pobreza, anual de 2006 a 2021; partidas de turistas, e PIB per capita anual, de 2006 a 2020; mortalidade por influenza em idosos, somente em 2011; disponibilidade de vacina para influenza, anual de 2018 a 2021; taxa de vacinação, anual de 2018 a 2021; número de pessoas que moram no país mas não nasceram no país, anual de 2010 a 2015; irregularidades em temperatura, mensal de 2006 a 2024; temperatura, mensal de 2006 a 2024; taxa de cobertura vacinal, anual de 2006 a 2018. A estação do ano foi definida de acordo com as datas mensais disponíveis no número de casos.

Os classificadores regressão logística, random forest e SVM foram implementados considerando como desfecho as variáveis que indicavam ou não a ocorrência de uma variação mensal nas taxas superior a 60%, 70% ou 80%. As medidas de desempenho ponderadas (acurácia, precisão, revocação e *score* F1) foram estimadas pela média da validação cruzada (CV) estratificada 10 vezes, repetida três vezes. Em cada iteração, os algoritmos foram treinados em nove partições e avaliados na partição restante. Consideramos a CV aninhada cinco vezes para o ajuste de hiper parâmetros, mais apropriada para uma avaliação confiável das medidas de desempenho. Os procedimentos de imputação, pré-processamento (padronização de variáveis numéricas e codificação one-hot para variáveis categóricas), necessários para garantir a qualidade e a consistência dos dados e para criar um classificador confiável e robusto, foram executados dentro da CV, nas partições de teste e treinamento separadamente, para evitar viés na estimativa de desempenho.

O classificador random forest apresentou melhor desempenho com medidas de acurácia, revocação e precisão acima de 0,70. Dentre as variaveis que apresentaram maior importância para a predição estão as variáveis identificadoras dos países Brasil, Peru e Paraguai, o número de partidas de turistas, número de pessoas que moram no país mas não nasceram no país e o ano.

O modelo construído demonstrou potencial para contribuir com sistemas de vigilância de influenza na América Latina sugerindo um aumento na sensibilidade do sistema (capacidade de detectar epidemias) e na oportunidade de detecção (velocidade para o desenvolvimento das etapas da vigilância). O aprendizado de máquina é uma estratégia promissora para ampliar a capacidade de vigilância regional e predição de novas epidemias de influenza na América Latina.

Referências

[1] J. R. Spinardi, K. B. Thakkar, V.L. Welch et al. "The need for novel influenza vaccines in low- and middle-income countries: A narrative review". Em: **The Brazilian Journal of Infectious Diseases** 1 (2025), p. 104465. DOI: https://doi.org/10.1016/j.bjid.2024.104465.

[2] N. Velappan, A. R. Daughton, G. Fairchild et al. "Analytics for investigation of disease outbreaks: web-based analytics facilitating situational awareness in unfolding disease outbreaks". Em: **JMIR Public Health and Surveillance** 1 (2019), e12032. DOI: https://doi.org/10.2196/12032.

[3] WHO. **Managing epidemics: key facts about major deadly diseases**. World Health Organization, 2023. ISBN: 9789240083196.