Algoritmos de Regressão Aplicados à Predição de Casos de Arboviroses no Cariri Paraibano

José G. da Silva Lima[1], Pedro L. G. Prata[1], José G. L. Filho[1], Roberto R. C. de Franca[1], Tiago B. Araujo[1]

[1]Instituto Federal da Paraíba (IFPB)
Monteiro – PB – Brasil

{silva.jose, preto.prata}@academico.ifpb.edu.br
{jose.filho, roberto.franca, tiago.brasileiro}@ifpb.edu.br

Abstract. *The present study intends to collaborate with the understanding of the high incidence of arboviruses in Paraíba, investigating the Cariri region, using Artificial Intelligence (AI) techniques for a broad understanding of the problem. It also cooperates with national development in the area of Science and Technology, in the search for computational techniques that can assist in the construction and validation of actions to combat arboviruses transmitted by Aedes aegypti. Two machine learning algorithms were used to compare results: Random Forest Regressor and Linear Regression. From the results obtained, it is possible to evidence that the algorithms presented the same results, being, therefore, promising alternatives for the prediction of Dengue.*

Resumo. *O presente estudo pretende colaborar com a compreensão da elevada incidência de arboviroses na Paraíba, investigando a região do cariri, utilizando-se de técnicas de Inteligência Artificial (IA) para o amplo entendimento do problema. Também, coopera com o desenvolvimento nacional na área de Ciência e Tecnologia, na busca de técnicas computacionais que possam auxiliar na construção e validação de ações para o combate das arboviroses transmitas pelo Aedes aegypti. Foram utilizados dois algoritmos de aprendizagem de máquina para comparação de resultados: Random Forest Regression e Regressão Linear. A partir dos resultados obtidos, é possível evidenciar que os algoritmos apresentaram resultados similares, sendo, portanto, alternativas promissoras para a predição de casos de arboviroses.*

1. Introdução

Arboviroses são doenças causas pelos chamados arbovírus, que incluem o vírus da dengue, Zika, Febre Chikungunya e Febre Amarela. Os arbovírus englobam todos os vírus hospedados essencialmente por artrópodes, como mosquitos, e que hoje são uma das principais causas de doenças infecciosas por todo o mundo [Cardoso et al. 2015]. Dengue é uma das arboviroses transmitidas pelo mosquito *Aedes aegypti*, com milhões de casos nos últimos anos no Brasil e no mundo [Araujo 2019].

Dentre os artrópodes, o *Aedes aegypti* tem representado um grande problema à Saúde Pública através disseminação de doenças, como: Dengue, Zika, Chikungunya, entre outras. O controle desses vírus é um grande desafio, principalmente pela necessidade de ferramentas computacionais que visam realizar o monitoramento e controle a potencializar a busca ativa e passiva do mosquito [Morrison et al. 2004; Who 1997].

Em 2020, até a Semana Epidemiológica 36, a região Nordeste registrou 140.527 casos prováveis de Dengue, 48.084 casos de Chikungunya e 4.442 casos de Zika. O Nordeste lidera com as maiores taxas de incidência por habitante quando se trata da Chikungunya e Zika, o que corrobora com a carência de ferramentas de combate ao mosquito como controle químico, epidemiológico e tecnológico. Por esse motivo, o combate destas arboviroses é considerada uns dos grandes desafios de saúde pública enfrentado pelo Brasil, e por muitas partes do mundo [Brasil 2020].

A elevada incidência de doenças causadas pelo *Aedes aegypti* pode ser analisada e estudada com diversas técnicas, especificamente estatísticas e/ou computacionais, que podem ajudar os serviços de saúde no controle das epidemias, bem como no efetivo direcionamento de recursos financeiros, pessoais, entre outros [Pinheiro 2020].

Logo, a utilização de técnicas computacionais sob uma análise associativa das informações contidas nos boletins epidemiológicos, tendo o intuito de encontrar relações atípicas ou típicas, resumindo os dados em novas formas podem servir como ferramentas tanto compreensíveis quanto úteis para saúde pública. Tais métodos vêm sendo utilizados em diversas áreas, tendo em vista sua extensa aplicabilidade em resultados válidos, implementáveis e de grande importância na resolução de problemas.

Diante disso, este trabalho apresenta uma abordagem interdisciplinar, com integração entre Ciência e Tecnologia, com a articulação da Ciência Biológicas, Computação e Estatística. O estudo pretende aplicar algoritmos Inteligência Artificial no que diz respeito ao entendimento dos casos de arboviroses (Dengue, Zika e Chikungunya) no Cariri Paraibano, investigando a microrregião.

2. Metodologia

A seguir são apresentadas as etapas utilizadas na metodologia deste trabalho para realizar a predição de doenças transmitidas pelo mosquito Aedes aegypti no Cariri Paraibano utilizando algoritmos de aprendizagem de máquina:

a. Pré-processamento: Inicialmente, os casos confirmados de Dengue, Chikungunya e Zika foram coletados, especificando a característica de cada um desses casos. Em seguida, dados meteorológicos do Instituto Nacional de Meteorologia (INMET) também foram coletados. Após a coleta, os dados foram tratados, eliminando os ruídos e mantendo a homogeneização dos atributos.

b. Processamento: A predição de doenças padece do problema de tratar a deficiência na quantidade de dados disponíveis nas plataformas digitais do governo brasileiro. Para solucionar esse problema, os casos de arboviroses foram sumarizados dentro das semanas epidemiológicas definidas dentro da legislação. Com os dados tratados, aplica-se o algoritmo de aprendizagem de máquina de regressão, gerando modelos de aprendizagem para utilização na predição de novos casos.

c. Análise: A partir do modelo de aprendizagem gerado analisa-se os resultados em relação ao seu desempenho, sendo possível utilizá-lo na predição de novos casos de Dengue, Chikungunya e Zika.

3. Resultados

Para os fins da pesquisa, foi desenvolvida uma base de dados composta de variáveis de precipitação, pressão atmosférica, temperatura e umidade relativa do ar, providas pelo

INMET e os dados dos casos notificados das arboviroses dentro das semanas epidemiológicas da 5° Região de Saúde da Paraíba ou Cariri Ocidental (Amparo, Monteiro, Ouro Velho, Serra Branca, Sumé, Taperoá e etc), oriundos da Secretaria da Saúde do Estado da Paraíba, através do Sistema de Informação de Agravos de Notificação do Ministério da Saúde (SINAN/PB). Os dados utilizados correspondem ao período de 2012 a 2020.

Cabe ressaltar que os dados climáticos provindos do INMET foram sumarizados em média e desvio padrão da ocorrência da variável na semana epidemiológica correspondente. Já nos casos de Dengue, Zika e Chikungunya foram coletados considerando o número de casos de todos os municípios dentro das semanas epidemiológicas. As notificações utilizadas não tiveram critérios de seleção, ou seja, foi utilizado o número total de casos.

Inicialmente, dispunha-se de 7.812 casos de dengue classificados, 97 casos de Zika e nenhum de Chikungunya. Devido ao diminuto número de casos de Zika, Chikungunya e à falta de alguns atributos relevantes, optamos pela predição, exclusivamente, de casos de Dengue na presente pesquisa. Por fim, após o refinamento dos dados, a base de dados ficou composta de 34 variáveis climáticas e uma para predição de casos de arboviroses (Dengue), totalizando 35 atributos.

Antepondo as análises da aplicação dos algoritmos de aprendizado de máquina, foi aplicado o algoritmo *Principal Components Analysis* (PCA) (DUNTEMAN, 1989) sobre a base de dados com intuito de reduzir a dimensão dos dados, porém mantendo-se informações e características. Após aplicação do PCA, a base de dados sofreu uma substituição nos dados de entrada por dados equivalentes, mas com uma dimensão reduzida com 2 atributos.

Após aplicação do PCA, foram realizados experimentos computacionais com os seguintes algoritmos: *Random Forest Regression* (RFR) [Segal 2004] e Regressão Linear [Seber e Lee 2012]. As implementações utilizadas estão disponíveis no pacote *sklearn* do Python 3.9. O algoritmo RFR foi parametrizado através do uso da técnica de GridSearch. Essa técnica testou todas as possibilidades de combinações possíveis para os hyperparâmetros, exaustivamente. Após essa fase de ajustes, os parâmetros definidos foram: 1) números de estimadores igual a 100; e 2) Máxima Profundidade sem limitação. Já no algoritmo de Regressão Linear, foram utilizados os paramêtros *default* que o *skylearn* prover ao usuário.

Em seguida, foi conduzida a fase de teste da precisão dos algoritmos e o enviesamento de algoritmos de aprendizado supervisionado, foi escolhida a técnica de validação cruzada através do 10-fold. Nessa, os dados foram divididos em 5 subconjuntos de tamanho igual e mutuamente exclusivos. São feitas 10 iterações para gerar 10 modelos, sendo que em cada um, 9 são testados para predição no conjunto restante, ou seja, todos os conjuntos devem estar presentes em um dos modelos como subconjunto predito pelos demais. A acurácia é obtida através da mensuração dos erros encontrados.

Por fim, os resultados obtidos foram coletados e, posteriormente, comparados quanto a medida da Raiz do Erro Quadrático Médio (RMSE) [Chai e Draxler 2014], uma das principais métricas utilizadas para avaliação do desempenho de algoritmos de regressão. O RMSE é interpretado da seguinte forma, quanto mais distantes às predições foram do real, maior será seu valor. Durante essa etapa de processamentos, foram alcançados os seguintes resultados: 1) Random Forest – 74.62; e 2) Regressão Linear –

67.77. Os resultados mostram que os algoritmos obtiveram desempenhos similares levando em consideração o RMSE como medida. Essa métrica se refere a capacidade que o algoritmo tem de predizer corretamente os casos de Dengue, conforme os dados utilizados na validação do modelo.

4. Conclusão

Este trabalho justifica a afirmativa de que a distribuição de dados relacionados aos registros de casos de vítimas de arboviroses por parte das instituições de saúde, podem ser úteis para o processo de mineração de dados, bem como para outras tarefas de extração de conhecimento e tomada de decisão. Nota-se, ainda, a necessidade de melhoria da qualidade dos dados oferecidos pelos portais de dados abertos do governo brasileiro. Isso permitirá que em trabalhos futuros possamos promover métodos mais assertivos baseados em IA que auxiliem a gestões públicas e privadas no combate as arboviroses.

Uma extensão dos estudos realizados pode, além de Dengue e Chikungunya, considerar e predizer outras doenças transmitidas pelo vetor Aedes Aegypti. Além disso, se propõe a implementação de uma ferramenta tecnológica a ser disponibilizada à sociedade, de modo que cidadãos, profissionais da saúde e gestores públicos possam se beneficiar, permitindo a recomendação do diagnóstico dessas doenças.

Referências

ARAUJO, S. (2019). Dengue e seus avanços. Rev. bras. anal. clin, p. 196-201.

BRASIL, Ministério da Saúde. Boletim Epidemiológico 38. Brasília, set. 2020. Disponível em:<https://antigo.saude.gov.br/images/pdf/2020/September/24/Boletim-epidemiologico-SVS-38.pdf >. Acesso em: 02 março de 2021.

CARDOSO, B. F. et al. Detection of Oropouche virus segment S in patients andin Culex quinquefasciatus in the state of Mato Grosso, Brazil. Mem Inst. Oswaldo Cruz, [S.L], v. 110, n. 6, p. 745-54, 2015.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development*, *7*(3), 1247-1250.

Dunteman, G. H. (1989). *Principal components analysis*. Sage.

MORRISON, A. et al. (2004) Evaluation of a Sampling Methodology for Rapid Assessment of Aedes aegypti Infestation Levels in Iquitos, Peru. Journal of Medical Entomology, Peru, v. 41, n. 3, p. 502-510.

SEBER, George AF; LEE, Alan J. (2012). Linear regression analysis. John Wiley & Sons.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

WHO, World Health Organization. Dengue hemorrhagic fever: diagnosis, treatment, prevention and control. Geneva, 1997. Disponível em: http://www.who.int/sorry/en/. Acessado em: 03 mar. 2021.

PINHEIRO, T. F., ALVES, J. B., SILVA, Y. R. N. (2020). O impacto financeiro das arboviroses oriundas do Aedes Aegypti no Brasil: uma projeção para 2019 / The financial impact of arboviroses from Aedes Aegypti in Brazil: a projection for 2019. Brazilian Journal of Development, Curitiba, v. 6, n. 5, p. 30757-30767.