ISSN: 2319-0124
josif 2023
15ª JORNADA CIENTÍFICA E TECNOLÓGICA
12º SIMPÓSIO DE PÓS-GRADUAÇÃO
DO IFSULDEMINAS
REALIZAÇÃO INSTITUTO FEDERAL Sul de Minas Gerais
APOIO FAPEMIG CNPq

UTILIZAÇÃO DO ALGORITMO SUPPORT VECTOR MACHINE PARA PREVISÃO DE CASOS DE DENGUE

Alexandre Oliveira Marinho[1] ;Hiran Nonato Macedo Ferreira [2];

RESUMO

A dengue, endêmica em regiões tropicais devido ao Aedes aegypti, pode causar grande número de casos e mortes. Usando dados do InfoDengue, este estudo aplicou métodos de Análise Exploratória de Dados (AED) e mineração de dados, incluindo o algoritmo de aprendizado de máquina (ML) Support Vector Machine, para identificar atributos que influenciam os casos.

Palavras-chave:
Dengue; Data Mining; analise exploratória de dados; Machine Learning.

1. INTRODUÇÃO

A dengue é uma arbovirose[3] causada pelo agente etiológico do gênero Flavivírus, transmitido por mosquitos do gênero Aedes, com maiores ocorrências em países com áreas tropicais e subtropicais (FIRMINO; SOUSA, 2023, PAHO, 2023). Tende a ocorrer ao longo dos trópicos, com variações locais de risco influenciadas pela precipitação, temperatura e rápida urbanização não planejada. A dengue é considerada umas das arboviroses mais frequentes do mundo, tendo em média aproximadamente 400 mil casos por ano.

Como forma de contornar esses desafios, técnicas da área de Inteligência Artificial (IA) vêm sendo utilizadas para auxiliar na resolução de problemas e pesquisas nesse campo. No estudo realizado por Doni e Sasipraba (2020), foram empregadas técnicas de Deep Learning (DL) para analisar dados climáticos, como temperatura, precipitação e umidade. O trabalho obteve uma elevada eficácia ao prever casos de dengue, alcançando uma acurácia de 89%. Outro estudo, conduzido por Mussumeci e Coelho (2020), utilizou técnicas de Machine Learning (ML) utilizando dados extraídos do Info Dengue, um sistema integrado de alerta de dengue. Foram aplicados os algoritmos Long Short Term Memory (LSTM), Random Forest e Lasso em suas versões de regressão. De acordo com o trabalho, o algoritmo que alcançou uma melhor eficácia foi o LSTM.

Deste modo, mostra-se possível utilizar técnicas ML para automatizar a predição de surtos de casos de dengue utilizando dados climáticos. ML ou, em português, aprendizagem de máquina, é um subcampo da IA dedicado ao desenvolvimento de algoritmos e técnicas que permitam a máquina aprender, possibilitando que a máquina aperfeiçoe seu desempenho em alguma tarefa específica por meio da aprendizagem (REZENDE, 2003). Ligada à mineração de dados e a estatística, a área de ML

[1] Bolsista PIBIC/CNPq, IFSULDEMINAS – *Campus* Passos. E-mail: alexandreomar@gmail.com
[2] Orientador, IFSULDEMINAS – *Campus* Passos. E-mail: hiran.ferreira@ifsuldeminas.edu.br
[3] Arbovirose: doenças causadas por vírus transmitidos, principalmente, por mosquitos.

foca nas propriedades dos métodos estatísticos, assim como sua complexidade computacional (AMORIN et al., 2008).

Neste contexto, existem algoritmos cujo objetivo é utilizar o método de regressão para prever um atributo de valor contínuo associado a um objeto. Neste estudo foi observado se o algoritmo Support Vector Machine (SVM) será eficaz para prever as possibilidades de surtos de casos de dengue do estado de Minas Gerais.

2. MATERIAL E MÉTODOS

2.1 Base de dados

Neste trabalho, forão coletados dados do InfoDengue[4] por meio da API (Interface de Programação de Aplicações), que se baseia em dados híbridos gerados pela análise integrada de dados minerados da web social, bem como de dados climáticos e epidemiológicos.

Os dados são disponibilizados no formato de csv, json e xml. Neste caso os dados que foram coletados foram no formato csv, que tem os seguintes atributos:

data_iniSE	casos_est	casos_est_min	casos_est_max	casos	p_rt1	p_inc100k
Localidade_id	nivel	id	versao_modelo	tweet	Rt	pop
tempmin	umidmax	receptivo	transmissao	nivel_inc	umidmed	umidmin
tempmed	tempmax	casprov	casprov_est	casprov_est_min	casprov_est_max	casconf
notif_accum_year						

O dicionário de dados pode ser obtido no próprio site do InfoDengue disponível na url[5].

2.1. Ferramentas utilizadas

O processo de coleta de dados, seleção de dados, Análise Exploratória de Dados (AED) e Data Mining foram executados utilizando a linguagem Python, que de acordo com Borges (2014) é de fácil aprendizado e possui vasta literatura disponível. E como ambiente de programação Python, está sendo utilizada a IDE (Integrated Development Environment) Jupyter Notebook.

Na coleta e seleção de dados foi utilizado a API disponibilizada pelo InfoDengue utilizando

[4] InfoDengue é um sistema de alerta para arboviroses criado e desenvolvido por pesquisadores do Programa de Computação Científica (Fundação Oswaldo Cruz, RJ) e da Escola de Matemática Aplicada (Fundação Getúlio Vargas) com a forte colaboração da Secretaria Municipal de Saúde do Rio de Janeiro, o Observatório da Dengue/UFMG e pesquisadores da Universidade Federal do Paraná e da Universidade Estadual do Oeste do Paraná (INFODENGUE,2023).

[5] url: https://info.dengue.mat.br/services/api.

as Bibliotecas do Python: pandas, seaborn e csv. Foram coletadas informações do estado de Minas Gerais (MG) dos últimos 5 anos. Por meio dessa API, é utilizado um laço de repetição para coletar todos os dados de todas as cidades de MG por meio dos parâmetros (por meio de requisição):

- geocode:código IBGE da cidade
- disease: tipo de doença (str:dengue|chikungunya|zika)
- format: extensão/ formato do arquivo (str:json|csv)
- ew_start: consulta inicial da semana epidemiológica(int:1-53)
- ew_end: consulta final da semana epidemiológica(int:1-53)
- ey_start: ano de consulta inicial(int:0-9999)
- ey_end: ano de consulta final(int:0-9999)

A AED utilizará as bibliotecas pandas, seaborn e sklearn. A biblioteca Pandas foi empregada para importar, observar e manipular os dados do arquivo CSV. O seaborn foi utilizado para representar graficamente a matriz de correlação, otimizando os atributos para o Aprendizado de Máquina (AM). A biblioteca Sklearn será empregada para normalizar os dados.

Com o uso da biblioteca pandas, uma nova coluna denominada 'IBGE' foi adicionada para melhor representação das cidades. Foram removidos tipos de dados inadequados, incluindo as linhas duplicadas e atributos do tipo objeto: 'data_iniSE' e 'versao_modelo'. Colunas com grande quantidade de valores vazios, como 'umidmed', 'umidmin', 'tempmed', 'tempmax', 'casprov', 'casprov_est', 'casprov_est_min' e 'casprov_est_max', tiveram suas linhas vazias substituídas por 0. Outras colunas que continham valores vazios e permaneceram foram tratadas da mesma forma. Colunas que poderiam introduzir vieses indesejados, como 'notif_accum_year', 'p_rt1' e ‘nivel’, foram removidas.

O algoritmo escolhido, Support Vector Machine (SVM), baseia-se na teoria de aprendizado estatístico. O objetivo do algoritmo SVM é encontrar a melhor linha ou fronteira de decisão que pode separar um espaço n-dimensional em classes, permitindo que novos pontos de dados sejam classificados corretamente no futuro (KURANI, 2023).

A matriz de correlação foi utilizada para identificar atributos irrelevantes, refinando assim o algoritmo SVM. A normalização dos dados será realizada por meio da biblioteca Sklearn. A construção e avaliação do algoritmo de Aprendizado de Máquina (SVM) também serão executadas usando a biblioteca Sklearn. Isso envolverá a definição de atributos, o treinamento das classes e a realização de testes usando métricas de desempenho. Foram utilizadas as métricas 'score' do Sklearn, além do Erro Quadrático Médio (Mean Squared Error) e do Erro Quadrático Médio da Raiz (Root Mean Square Error - RMSE) da biblioteca numpy.

3. RESULTADOS E DISCUSSÃO

Após as etapas de seleção de dados, analise exploratória de dados e normatização, foi possível

testar e ter um resultado parcial do algoritmo Support Vector Machine Regression (SVR). Primeiramente foram definidas as variáveis/dados de treino e teste, com o tamanho do teste(test_size) em 0.3 e o estado de aleatoriedade em 100, também os parâmetros do algoritmo foram detidos como 'C': 100, 'gamma': 0.01, 'kernel': 'rbf'. Após a definição o algoritmo foi aplicado, podendo se obter estes resultados:

- Score: 0.7745680097196351
- MSE: 0.36
- RMSE: 0.6

4. CONCLUSÃO

Ao utilizar os dados de Minas Gerais obtidos pelo Info Dengue, foi possível observar que os atributos mais relevantes para a quantidade de casos no dataset são: p_inc100k, tweet, pop, receptivo, transmissão e nível_inc. Vale destacar que esses atributos também estão inter-relacionados.

Com os resultados obtidos, foi evidenciado que, ao utilizar os dados coletados, o algoritmo Support Vector Machine Regression (SVR), com todos os parâmetros definidos, obteve resultados considerados bons. Ele foi capaz de prever aproximadamente 77,4% dos dados.

REFERÊNCIAS

AMORIM, Maurício JV; BARONE, Dante; MANSUR, André Uebe. Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2008. p. 666-674.

DONI, Anjelus Ronald; SASIPRABA, Thankappan. LSTM-RNN Based Approach for Prediction of Dengue Cases in India. Ingénierie des Systèmes d'Information, v. 25, n. 3, 2020.

FIRMINO, Luan Cesar Correia; DE SOUSA SOUSA, Milena Nunes Alves. Educação em Saúde como Estratégia de Enfrentamento da Dengue: Um Relato de Experiência. ID on line. Revista de psicologia, v. 17, n. 65, p. 313-322, 2023.

KURANI, Akshit et al. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, v. 10, n. 1, p. 183-208, 2023.

MUSSUMECI, Elisa; COELHO, Flávio Codeço. Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression. Spatial and Spatio-temporal Epidemiology, v. 35, p. 100372, 2020.

Paho. Dengue. Disponível em :<https://www.paho.org/pt/topicos/dengue>. Acesso em: 12 de março de 2023.

REZENDE, Solange Oliveira. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.