DengAI

Predicting Disease Spread

Dima Mamdouh Mohamed
2110000811
D.Mamdouh2181@nu.edu.eg

Youssef Hassan Abdelmaksoud
211002056
Y.hassan2156@nu.edu.eg

Ahmed Kamal Ahmed
211000202
A.kamal2102@nu.edu.eg

Omar Ayman Morshdy
211001749
O.*ayman2149@nu.edu.eg*

Mohanad Ashraf Abdelwahab
222000061
m.ashraf2261@nu.edu.eg

Mohamed Galal Elgemeie
202000206
M.elgemeie@nu.edu.eg

***Abstract*— Dengue fever, a widespread mosquito-borne disease prevalent in tropical regions, poses a significant global health threat. This research aligns with the Predict the Next Pandemic Initiative, leveraging collaborative data from institutions like the U.S. Centers for Disease Control and Prevention. Using machine learning models and data scraping techniques, we forecast Dengue fever incidence in San Juan, Puerto Rico, and Iquitos, Peru, contributing to global efforts in predicting and preventing pandemics. The methodology employs seven machine learning models, addressing complexities associated with emerging infectious diseases. Results showcase the efficacy of models, emphasizing the need for tailored approaches in disease prediction. The discussion interprets findings, guiding researchers and practitioners in selecting models for accurate Dengue forecasting.**

Keywords— Dengue fever, machine learning, predictive modeling, infectious diseases, climate variables, public health, pandemic prediction, data scraping

I. Introduction

Dengue fever, a mosquito-borne illness prevalent in tropical regions, poses a significant global health threat with symptoms ranging from flu-like discomfort to severe complications. The disease's transmission intricately links to climate variables, emphasizing the need for adaptive strategies in the face of an evolving threat. This research aligns with the Predict the Next Pandemic Initiative, involving key institutions such as the U.S. Centers for Disease Control and Prevention. Leveraging collaborative data, it provides a nuanced understanding of the interplay between environmental factors and Dengue incidence, contributing to ongoing efforts in predicting and preventing future pandemics. Furthermore, this paper outlines a methodology integrating machine learning models and data scraping techniques to forecast Dengue fever incidence in San Juan, Puerto Rico, and Iquitos, Peru. Focused on advanced analytics, it addresses complexities associated with emerging infectious diseases, emphasizing interpretability, and visualization for valuable insights. The outcomes are expected to contribute to global efforts in combatting infectious diseases.

II. Related Works

Several studies have delved into the relationship between climate conditions and the incidence of Dengue fever, recognizing the significance of predictive modeling in understanding and managing this public health challenge. Baker et al. [5] focused on forecasting Dengue Fever using machine learning regression techniques. Their work contributes to the growing body of research aiming to predict Dengue cases based on climate variables, aligning with the idea that improved outbreak prediction can aid in disease prevention and control.

In the broader context of predictive analytics for disease analytics, Souza et al. [6] proposed an innovative big data predictive analytics framework over hybrid big data sources, showcasing an application for disease analytics. While not specifically focused on Dengue, their work provides insights into leveraging advanced analytics techniques for disease prediction, which can be relevant to our study.

These studies underscore the importance of employing predictive models to anticipate Dengue outbreaks based on climate data. However, our research distinguishes itself by specifically applying and comparing various machine learning regression algorithms to predict Dengue cases in the cities of San Juan and Iquitos, providing a nuanced analysis of the most effective modeling approach.

III. Methodology

This paper outlines a comprehensive methodology for predicting Dengue fever incidence, leveraging a diverse set of seven machine learning models. The methodology encompasses the following key stages:

A. Data Pre-Processing

A. Cleaning and Visualization: The initial step involves thorough data cleaning to address any inconsistencies or missing values. Visualization techniques are applied to gain insights into the dataset's distribution and relationships.

B. .Data Splitting and Cross-Validation:

- The dataset is split into training and validation sets to facilitate model training and evaluation. Cross-validation techniques ensure robust model generalization and performance assessment on unseen data.

979-8-3503-4726-5/23/$31.00 ©2023 IEEE

C. Performance Evaluation

Rigorous performance metrics, such as Mean Absolute Error (MAE), are employed to assess the predictive capabilities of each model.

D. Ensemble Modeling:

An ensemble approach was adopted to combine the predictive power of multiple machine learning models. This technique enhanced overall model accuracy and reliability by leveraging the diversity of individual models.

E. Hyperparameter Tuning:

Model hyperparameters were fine-tuned to optimize predictive performance. A systematic exploration of hyperparameter space identified configurations yielding the most effective models

F. Machine Learning Models

1) Linear Regression (LR_model):

   The Linear Regression model is employed as a foundational tool to capture linear relationships within the dataset. Specifically, it serves to model the impact of linear environmental variables on Dengue fever incidence. For instance, it can be applied to establish a clear correlation between temperature variations and the occurrence of Dengue cases.

2) Support Vector Machine (SVM) Model:

   The Support Vector Machine (SVM) model plays a pivotal role in handling high-dimensional data and discerning complex, nonlinear correlations. In the context of this research, SVMs are applied to establish robust decision boundaries, particularly in navigating the intricate landscape of climate variables. By doing so, SVMs contribute significantly to the accuracy of Dengue incidence predictions.

3) Decision Tree Model:

   Decision Trees are integrated into the methodology for their transparency and interpretability. They unfold the decision-making process and reveal influential environmental factors affecting Dengue fever occurrence. For example, Decision Trees can identify critical features such as precipitation levels, providing valuable insights into Dengue transmission dynamics.

4) Random Forest Model:

   The Random Forest model is employed to address overfitting concerns and capture complex relationships through ensemble learning. By aggregating insights from multiple Decision Trees, Random Forests enhance the robustness of Dengue incidence prediction. This ensemble approach offers a comprehensive understanding of various climate factors influencing Dengue patterns.

5) Gaussian Naive Bayes Model:

   Naive Bayes is incorporated into the methodology for its computational efficiency and effectiveness in probabilistic classification. This model calculates the probability distribution of Dengue incidence based on environmental features. For instance, Naive Bayes can be applied to estimate the likelihood of Dengue outbreaks under specific climate conditions, offering probabilistic insights.

6) K-Nearest Neighbors (KNN) Model:

   K-Nearest Neighbors (KNN) is utilized for its simplicity and effectiveness in capturing local patterns within the dataset. In this research, KNN aids in identifying regions with similar Dengue incidence patterns, contributing to a spatial understanding of the disease. For example, KNN can pinpoint areas with analogous transmission dynamics, providing valuable spatial insights.

7) Neural Network (MLP) Model:

   The Neural Network, specifically the Multi-layer Perceptron (MLP) model, is incorporated to capture intricate, non-linear relationships in high-dimensional data. It is applied to understanding complex interactions between climate variables impacting Dengue fever. For example, MLPs unveil nuanced relationships between various climate factors and Dengue incidence, facilitating a sophisticated analysis of disease dynamics.

IV. Results

In evaluating the predictive models for Dengue cases based on climate data, key performance metrics were employed to assess their accuracy and reliability. The Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (r2) were scrutinized for each model. The Linear Regression model demonstrated a MAE of 10.07, RMSE of 13.84, and an r2 of 0.34, indicating a moderate predictive capability. Notably, the Support Vector Machine (SVM) exhibited a lower MAE of 9.65, suggesting improved accuracy in predicting Dengue cases. The Decision Tree Regressor also performed well with a MAE of 9.70. However, challenges emerged with the Gaussian Naive Bayes model, showing a higher MAE of 16.75 and a negative r2, indicating limitations in capturing the non-linear relationships within the data. Overall, these results offer valuable insights for selecting models with superior predictive capabilities for Dengue forecasting.

V. Discussion

Interpreting the results of our predictive models unveils nuanced insights into their effectiveness and limitations in forecasting Dengue cases based on climate variables. The Linear Regression model, serving as a baseline, demonstrates a moderate predictive capability, emphasizing linear relationships within the data. The Support Vector Machine (SVM) outperforms other models, showcasing its robustness in handling complex, non-linear patterns, and confirming its suitability for Dengue prediction tasks.

The Decision Tree Regressor also exhibits commendable performance, indicating its ability to capture intricate relationships within the dataset. However, challenges arise with the Gaussian Naive Bayes model, which struggles to accommodate the non-linear dynamics inherent in Dengue transmission. This emphasizes the significance of selecting models capable of addressing the complex interplay of climate variables.

The outcomes underscore the importance of considering the inherent complexities of Dengue dynamics when choosing predictive models. The negative r2 for Gaussian Naive Bayes highlights its limitations in capturing the non-linear nature of the relationship between climate variables and Dengue incidence. While Linear Regression provides a foundational understanding, more sophisticated models like SVM and Decision Tree Regressor prove pivotal in capturing intricate patterns for improved prediction.

These findings contribute to the ongoing discourse on the applicability of machine learning in disease prediction, particularly for vector-borne diseases such as Dengue. The discussion encourages thoughtful consideration of model selection, highlighting the need for tailored approaches to accommodate the diverse and dynamic nature of disease transmission. Overall, our study provides valuable insights for researchers and public health practitioners, guiding them in selecting suitable models for accurate Dengue forecasting and proactive disease management.

REFERENCES

[1] DrivenData, "DengAI: Predicting Disease Spread," [Online]. Available: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/page/82/

[2] DrivenData, "Dengue Forecasting Benchmark," [Online]. Available: https://drivendata.co/blog/dengue-benchmark/ K. Elissa

[3] International Journal of Science and Research (IJSR), ISSN: 2319-7064, ResearchGate Impact Factor (2018): 0.28 | SJIF (2018): 7.426.

[4] Professor Joanne Luciano, "I590 – Data Science for Drug Discovery, Health and Translational Medicine," December 10, 2017.

[5] Q. B. Baker, D. Faraj and A. Alguzo, "Forecasting Dengue Fever Using Machine Learning Regression Techniques," 2021 12th International Conference on Information and Communication Systems (ICICS), Valencia, Spain, 2021, pp. 157-163, doi: 10.1109/ICICS52457.2021.9464619.

[6] Souza, J., Leung, C.K., Cuzzocrea, A. (2020). "An Innovative Big Data Predictive Analytics Framework over Hybrid Big Data Sources with an Application for Disease Analytics." In: Barolli, L., Amato, F., Moscato, F., Enokido, T., Takizawa, M. (eds) Advanced Information Networking and Applications. AINA 2020. Advances in Intelligent Systems and Computing, vol 1151. Springer, Cham. https://doi.org/10.1007/978-3-030-44041-1_59.