Hotspots Analysis of Malaria Prevalence

Use Case: Nigeria

By Sam Ekong

Project Overview

This project aims to identify and analyze malaria hotspots prevalence in Nigeria, leveraging geospatial data and epidemiological insights to inform targeted interventions and resource allocation. The project findings and recommendations are expected to contribute to reducing the burden of malaria in Nigeria, particularly in vulnerable populations, and improving overall public health outcomes.

Goals & Objectives

To identify hotspots, analyze malaria prevalence and its correlations, that can ease the disease burden in Nigeria; and to inform targeted interventions through spatial mapping of high risk areas, statistical EDA and trend factors contributing to malaria transmission, which seek to support evidence-based decision-making for malaria control strategies and recommendations for public health actions in Nigeria.

Description and Problem Definition

Malaria remains a significant public health challenge in Nigeria, with certain regions experiencing disproportionately high prevalence rates, underscoring the need for targeted and data-driven approach to effectively control and eliminate this disease to a reasonable scale.

Questions and Assertions

Research Questions

1. What are the primary factors contributing to the prevalence of malaria hotspots cases in Nigeria?
2. How can geospatial data and mapping techniques be leveraged to enhanced malaria surveillance and control in high-risk areas?
3. What interventions have been proven effective in reducing malaria prevalence in similar hotspots globally, and how can they be adapted for Nigeria?

Assertions

1. Targeted intervention in malaria hotspots zones can significantly reduce overall malaria prevalence and transmission rates in Nigeria.
2. Geospatial data-driven mapping are critical tools for identifying and characterizing malaria hotspots, enabling more effective resource allocation and intervention strategies.
3. Addressing the socioeconomic and environmental determinants of malaria is essential for achieving sustainable reduction in malaria prevalence in Nigerian hotspots areas.

Methodology

This project will employ a retrospective spatial analysis of malaria survey, data collection, preprocessing, geospatial preparation, visualization, spatial autocorrelation analysis combining it with environmental, demographic to identify hotspot mapping and model build to accurately predict statistically significant cases of malaria prevalence among other variables.

Data Resources and Tools

Grid3_Admin Boundary (NGA_State_Boundaries), Nigeria Malaria Survey (NMIS) from Ministry of Health, Nigeria Cities Weather Data by GeoBoundaries data provided under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://www.geoboundaries.org/), National Malaria Elimination Programme (NMEP), World Health Organization Surveillance datasets, Nigeria Open Disease & Health Data, Nigeria Malaria Indicator Survey, DHS Program (https://data.humdata.org), (https://pmc.ncbi.nlm.nih.gov), (https://www.geoboundaries.org). Tools like QGIS, cloud computing using Google Colab (Python).

Analysis and Techniques

- Spatial autocorrelation analysis is carried out combining it with environmental, demographic to identify hotspots.
- The analysis in these project involve spatial regression, modeled to examine the relationships between malaria prevalence and potential risk factors.
- Techniques such as spatial interpolation (KDE) and hotspot (Heatmap) analysis applied to visualize and quantify the distribution of malaria cases.

Discussion / Summary

The strong correlation between malaria prevalence and climatic factors aligns with existing analysis on malaria ecology. Warm, humid environments foster mosquito breeding and parasite development. The negative relationship with sea level suggests that low-lying areas are more prone to malaria transmission. Population density amplifies malaria risk, as urban centers may experience higher exposure rates despite varying levels of intervention. There is a performance improvements observed across the models after fine-tuning and scaling. Identifying the best-performing model based on the evaluation metrics leads the discussion for the significance of the results for malaria prevalence hotspot analysis, suggesting potential next steps for further model enhancement or real-world deployment. In view of the essence for this project and the compelling questions and answers, the data-driven insights are as follows:

Result Findings

Correlation Among Malaria Indicators

- Malaria prevalence correlated with RDT confirmation (0.56) and ACT confirmation (0.45).
- State-level aggregated malaria prevalence strongly correlated with mean RDT (0.80) and ACT (0.66).

Environmental Factors

Malaria prevalence strongly correlated with:

- Temperature (0.92)
- Humidity (0.81)
- Precipitation/Weather degree (0.95)
- Cloud coverage (0.90)
- Negative correlation with sea level (-0.97)

Population Factors

- Population size moderately correlated with malaria prevalence (0.56) and environmental variables.

Spatial Distribution

- Choropleth maps showed malaria prevalence concentrated in states with high humidity, temperature, and dense population clusters.

Performance Improvements After Fine-tuning and Scaling

- **Random Forest:** Initial test accuracy was 67.66%, improving to 71.14% after tuning.

- **Logistic Regression:** Initial test accuracy was 70.17%, changing slightly to 70.11% after tuning, but with crucial insights into its performance on the minority class.
- **XGBoost:** Initial test accuracy was 69.72%, improving to 71.27% after tuning.
- **Naive Bayes:** Test accuracy remained 62.43% after tuning, as GaussianNB has no primary hyperparameters for GridSearchCV to optimize in this context.

Best-Performing Model and Significance

Based on the evaluation metrics, **XGBoost** is the best-performing model, achieving the highest accuracy of 71.27% and demonstrating a relatively balanced performance across classes (recall for class 0 at 0.20, and for class 1 at 0.93). This makes it the most suitable model for identifying malaria prevalence hotspots, as it offers a better capability to detect the minority class (hotspots) compared to Logistic Regression, which completely failed to identify any instances of class 0.

Data Analysis Key Findings

- Data preprocessing involved handling missing values, column renaming, and applying MinMaxScaler to numerical features before a train-test split.
- Initial model training and evaluation showed varying performances: Naive Bayes (62.43% test accuracy), Random Forest (67.66% test accuracy), Logistic Regression (70.17% test accuracy), and XGBoost (69.72% test accuracy).
- Hyperparameter tuning using GridSearchCV led to performance adjustments for the models:
    - **Naive Bayes:** Maintained a test accuracy of 62.43% and an F1-score of 0.721 for class 1.
    - **Random Forest:** Achieved a test accuracy of 71.14% and an F1-score of 0.822 for class 1, with best parameters max_depth: 10, min_samples_split: 10, n_estimators: 100.

    - **Logistic Regression:** Achieved a test accuracy of 70.11% and an F1-score of 0.824 for class 1, with best parameters C: 0.01, solver: liblinear. Critically, it had 0% recall for the minority class (class 0), indicating it failed to identify any instances of malaria hotspots.
    - **XGBoost:** Achieved the highest test accuracy of 71.27% and an F1-score of 0.820 for class 1, with best parameters learning_rate: 0.2, max_depth: 3, n_estimators: 50, subsample: 0.8. It showed a better balance with a class 0 recall of 0.20 and class 1 recall of 0.93.
- This comparative analysis confirmed XGBoost as the top-performing model in terms of overall accuracy and balanced performance, making it the most suitable for detecting malaria prevalence hotspots.

Insights

- To improve the detection of malaria hotspots, priorities to addressing the class imbalance issue in the dataset, especially given Logistic Regression's complete failure to identify the minority class and XGBoost's still being relatively low on recall for class 0.
- Beyond simple accuracy, to also focus on metrics like recall and F1-score for the minority class (malaria hotspots) during further or subsequent model refinement and selection, as correctly identifying hotspots is crucial for intervention.

Conclusion

This project demonstrates the following:

- Malaria prevalence in Nigeria is influenced by environmental and demographic conditions.
- Targeted interventions in hotspot states, through the use of predictive risk modeling, and integrating multi-sectoral strategies.
- This comparative analysis confirmed XGBoost as the top-performing model in terms of overall accuracy and balanced performance, making it

the most suitable for detecting malaria prevalence hotspots.

Recommendations

- Target malaria control programs in states with high climatic suitability for malaria transmission (High-risk state require targeted interventions: bed nets, IRS spraying, awareness campaigns).
- Deploy geospatial risk maps to guide national malaria elimination campaigns.
- Integrate machine learning models for predictive risk assessment to support early warning systems.
- Encourage multi-sectoral approaches combining health, environmental, and urban planning data.