Building a Malaria Incidence Rate Predictive Model to Enhance the Understanding of Malaria Control Strategies

Chinenye Chukwu-Mba
Busayomi Thompson-Ajayi
Oloruntoba Gabriel Irojah
Deborah Popoola
Abednego Aginam
Gbadegesin Obaloluwa
Dike Calista
Okebiorun Omolola
Emmanuel Mugabo
Mayukh Chakraborti
Muhammad Asif
Olatunbosun Victor
Kandasamy K

1. INTRODUCTION

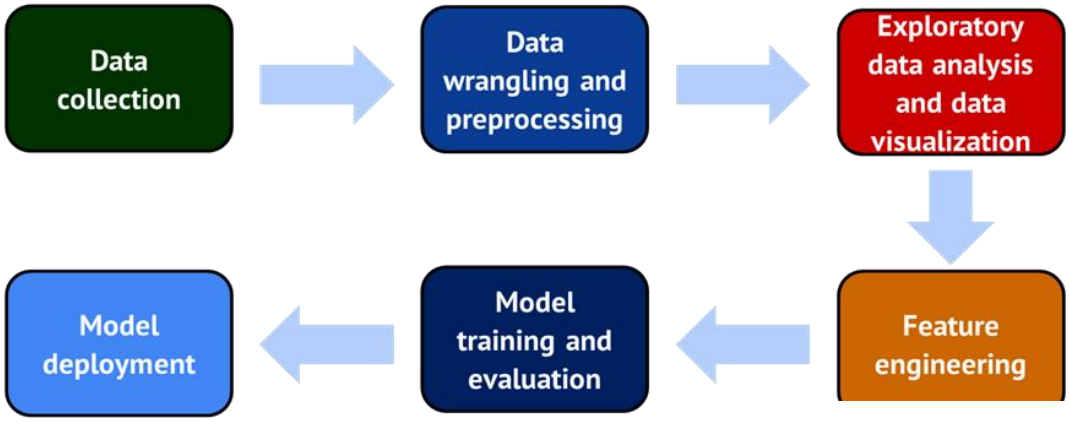
WHO's Global Technical Strategy for malaria has highlighted the importance of malaria surveillance as the third pillar for moving closer to malaria elimination. Effective surveillance data will assist countries in monitoring progress towards malaria elimination and targeting interventions to the last remaining at-risk places. An evaluation of the performance of a malaria incidence prediction model, created by Team Flask of Hamoye Winter Cohort 2023 was conducted, to identify key gaps which could be addressed to build effective systems for malaria elimination. Team Flask used machine learning techniques to facilitate the prediction of malaria incidence rates in Africa. We identified errors in their data handling, modeling techniques and inclusion of critical factors that affect malaria dynamics. To tackle these limitations, we (Team Gitlab) included new measurement indices and introduced new perspectives to their malaria prediction model. We also studied the possible benefits of interactions among existing malaria preventive measures. It is hoped that the techniques and models created through this study will be better equipped to guide the targeting of future interventions toward areas, populations, and sub-populations that are most ravaged by this disease.

2. AIM AND OBJECTIVES

This research aims to address critical lapses in malaria research done by Team Flask. Our research objectives include

- To assess the intricate interplay of population dynamics and malaria incidence rates, to discern the potential of incorporating population data as a predictive feature to amplify model generalization and predictive accuracy.
- To study the interaction amongst malaria preventive measures and identify complementary and region-specific measures.
- To create an improved malaria prediction model by engineering new features using our findings, introducing incidence rate thresholds and considering other modeling options based on research findings.

3. FLOW PROCESS



4. METHODOLOGY

Clinical Data:

The confirmed malaria incidence and annual population for 10 years ranging from 2007 to 2017 for all the thirty-six selected countries, was obtained from the World Bank data repository. The dataset contains a normalized value of the annual confirmed malaria incidence per 1000 population, which is the annual rate computed by dividing confirmed malaria incidence by its population size. The use of malaria control measures for 10 years, also ranging from 2007 to 2017 for all the thirty-six selected countries was obtained from both UNICEF child health coverage and World Health Organisation data repositories.

Data cleaning and pre-processing:

Machine learning principles uphold high accuracy through data preprocessing to obtain high predictive performance. Upon careful examination of the original dataset, it was observed that a significant portion of relevant have missing values. To solve this problem, we judiciously imputed these values using additional datasets obtained to fill in the missing values instead of discarding these columns or filling them with zeros. The additional datasets obtained from diverse sources were merged coherently and logically using country names and year as common columns to prevent the further occurrence of missing values. Some of the datasets added as new features included the estimated number of malaria deaths, estimated number of malaria confirmed cases, total population, rural population and urban population dataset. Each of these independent datasets also underwent specific data preprocessing. The data frame was also converted to a geo-dataframe to use features relating to the spatial information such as geometry, longitude and latitude. This conversion was carried out using Geopandas

Data Analysis and Feature Engineering:

Exhaustive data analysis was done alongside feature engineering to derive extensive insights into the analysis. In the process of feature engineering, we combined different aspects of the datasets to create new features. These new features included things like total malaria cases, standardized incidence rate, mortality rate, prevalence rate, and case fatality rate. The analysis conducted is outlined below

- A factor analysis of malaria indices (as variables) and malaria incidence report was carried out to examine the correlation and possible causal relationship between these variables and malaria

incidence. The combined impact of these variables on malaria incidence was observed. To do this, Bartlett's test of sphericity was used to determine the chi-square and P-values of our data. Also, model calculation was done and a scree plot of the eigenvalues obtained was created.

- Individual engineered features were analysed in terms of temporality and spatiality to draw further insights (please refer to the research article)
- A counterfactual scenario of the preventive measures to predict the incidence rates was modelled
- Spatial analysis was also performed using diverse techniques such clustering analysis, spatial autocorrelation and spatial regression using spreg module. In the spatial autocorrelation, two methods were used, this include
    - Global Moran's I's statistics
    - Hotspot / cold spots analysis using Local Getis-Ord Statistics (Gi)*

  Due to the limitation of the first method, the second approach was carried to investigate for the spatial pattern of malaria incidence rates.

Data Visualization:

Several corresponding data visualization was carried out. Some of the visualization done included interactive maps using folium to visually represent the malaria incidence rates by country and year. Geographical visualizations of malaria hotspot areas with respect to the hotspot analysis stated above were created. Temporal trends of malaria metrics i.e. incidence rates, prevalence rate, mortality rate, case fatality rates etc. were created. Most of these visualizations are included in the research article and presentation slides based on how we intend to use them.

Feature Selection:

The permutation feature importance techniques using Random Forest and XGBOOST guided in the selection of the features for modeling. Using the RMSE as error metrics, this technique evaluates the importance of each feature by permuting its values while keeping other features unchanged and then measuring the resulting decrease in model performance. Overall, from the selected features, only 5 of them were selected for easy interpretation and deployment. The selected features included 'Malaria cases reported', 'Malaria death', 'Total Malaria Cases', 'Total Population'

Model Building

In this project stage, four(4) models were built.

- Random Forest
- XGBOOST
- Stacking regressor using Random Forest as the final estimator
- Voting regressor

Model Training

The model training followed the conventional steps of standardizing the input variables using StandardScaler from the sklearn library, splitting the data into training and test sets, initializing the model, fitting the model, making predictions on test sets, evaluation using mean absolute error, mean squared error, root mean squared error and R^2 score.

Model Validation and Hyperparameter Tuning

A 10-fold cross-validation was performed on each training set for the baseline models. The mean and the standard deviation were also computed. For the hyperparameter tuning, a random search CV was employed factoring all the hyperparameter present in each model used.

Model Evaluation

During the evaluation, a number of metrics were considered and these included the RMSE and R^2 score. The model that has the lowest RMSE and high R^2 score was chosen and this was the Random Forest Regressor with an RMSE of 48 and R^2 score of 0.9918

Model Interpretability

The interpretability part was handled using LIME and then incorporating a baseline classification threshold to classify the predicted incidence rates into either low, medium or high.

Model Deployment

The best-performing model ie the Random Forest was selected and deployed on stream lit

5. RESULT

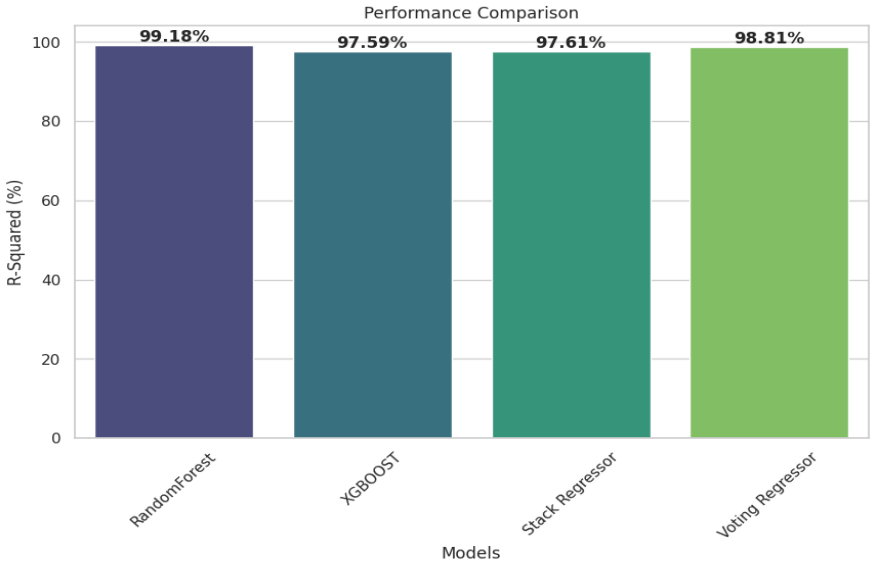
Below, we present the predictive performance of the models. The table provides an overview of the RMSE and R^2 scores achieved during the training and testing phases

Table 1: Baseline model performance on both train and test sets

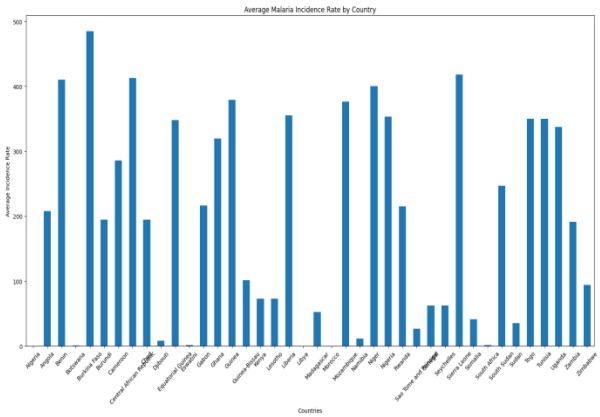
	RANDOM FOREST	XGBOOST	STACKING REGRESSOR	VOTING REGRESSOR
EVALUATION ON TRAINING SET				
RMSE	43.4975	0.0999	12.8840	21.7508
R2 SCORE	0.9939	0.999	0.9936	0.9985
EVALUATION ON TEST SET				
RMSE	44.4683	54.8749	42.8703	45.6289
R2 SCORE	0.9931	0.9895	0.9342	0.9927

Table 2: Model performance after cross-validation and hyperparameter tuning

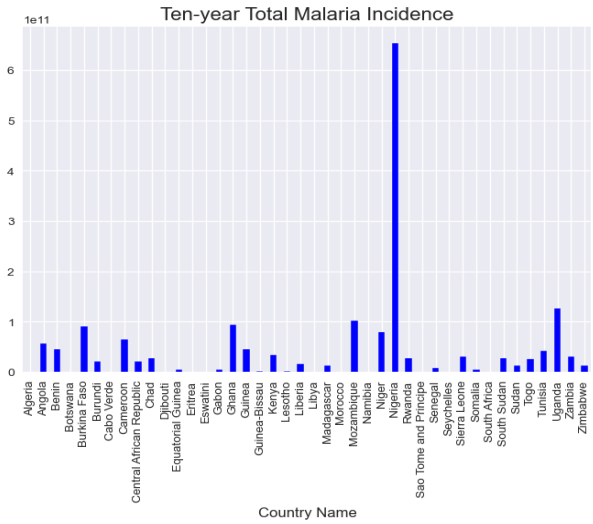
	RANDOM FOREST	XGBOOST	STACKING REGRESSOR	VOTING REGRESSOR
CROSS-VALIDATION EVALUATION ON TRAIN SETS				
RMSE	96.8500 ± 72.665	80.2247 ± 51.539	90.2765 ± 72.718	80.3541 ± 59.627
R2 SCORE	0.9144 ± 0.113	0.9574 ± 0.00345	0.9332±0.0765`	0.9502 ± 0.0604
HYPERPARAMETER TUNING – EVALUATION ON TEST SETS				
RMSE	48.3735	83.0140	82.6712	58.3468
R2 SCORE	0.9918	0.9759	0.9761	0.9881



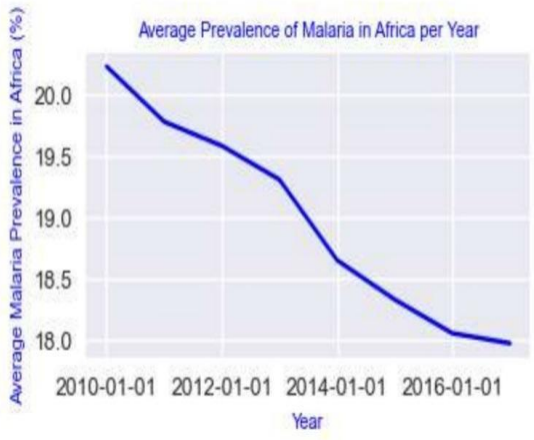
Performance comparison between different models



A bar graph of malaria incidence rates per 1000 population



A bar graph of total malaria incidence



Graph showing annual average of malaria prevalence in Africa between 2007 and 2017

6. CONCLUSION

This project presented the use of real-world data to classify malaria incidence in Africa, analyzing the same data for insights into the effectiveness of malaria control measures, selecting these features and building an improved malaria Incidence prediction model. The results suggest that the principal variable that influences malaria incidence varies from one country to another in different ways. Our dataset contained data only on malaria incidence in thirty-six African countries between 2007 and 2017. Future work can replicate and extend our work to other countries in Africa, as well as other countries where malaria is prevalent in the world. It can also make use of more current data. The main objective of this study was not only to locate African countries at a higher risk of malaria, but to build a more comprehensive computational model capable of predicting malaria incidence in Africa. For this project, we only considered variables with a

factor loading of at least 0.5. Consequently, the relationship between other variables and malaria incidence was not evaluated. In addition to the variables we have considered currently, seasonal changes like rainfall, the distribution of malaria parasites of varying types and other disease control programs, such as COVID-19, affect malaria control and resource management. Future work might consider these other factors considered to be contributing to malaria incidence as well as using other factor analysis methods to get the whole picture. They can consider further sophistication of the model to include these factors as features.

Team members

Chinenye Chukwu-Mba – Group Lead
Mayukh Chakraborti – Assistant Group Lead
Gbadegesin Obaloluwa David – Query Analyst
Busayomi Thompson-Ajayi
Dike Calista
Oloruntoba Gabriel Irojah
Okebiorun Omolola
Olatunbosun Victor
Deborah Popoola
Emmanuel Mugabo
Kandasamy K
Abednego Aginam