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

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

ABSTRACT

This research presents a comprehensive exploration of malaria incidence rates in Africa, aiming to enhance our understanding of control strategies. Leveraging data from diverse sources, including WHO and the World Bank, meticulous data preparation, feature engineering and advanced analytics were conducted. The analysis unearthed several critical insights: Temporal Trends: Malaria incidence and mortality rates exhibited a declining trend over the study period from 2007 to 2017. However, disparities among countries persisted, emphasizing the need for tailored interventions. Spatial Patterns: Non-random spatial patterns in incidence rates challenged the assumption of random disease dispersion. Clustering analysis identified hotspots, informing targeted resource allocation and interventions. Malaria metrics: the study introduced critical metrics such as Standardized Incidence Rates, Mortaltiy Rates and Case Fatality Rates, providing a nuanced perspective on malaria dynamics.

The development of predictive models, including Linear regressions, XGBOOST, Random Forest, Stacking Regressor and Voting Regressor with XGBOOST being the best performing model with an impressive R^2 score of 0.98 demonstrated the potential of data-driven decision support in malaria control strategies.

This research not only advances our understanding of malaria dynamics but also provides practical implication for public health. It guides the development of more effective control strategies and informs resource allocation, contributing to the ongoing fight against malaria.

INTRODUCTION

Malaria is a life-threatening disease primarily spread to humans by infective female anopheles mosquitoes. Female anopheles mosquitoes get infected by taking a blood meal from an infected human. Transmission between humans occur when plasmodium parasite pass through the saliva of a female anopheles mosquito into a person's bloodstream. Malaria may also be spread by transfusion of blood from infected people or by the use of contaminated parenteral drug administration equipment [1].

Symptoms, travel history, and a physical exam can cause a health care provider to suspect malaria. Laboratory tests, which show if a malaria parasite is present or not, will confirm a diagnosis [2]. Malaria parasites can be identified by examining under the microscope, a drop of the patient's blood, spread out as a blood smear on a microscope slide. Prior to examination, the specimen is stained (most often with the Giemsa stain) to give the parasites a distinctive appearance. This technique remains the gold standard for laboratory confirmation of malaria. However, it depends on the quality of the reagents, of the microscope, and on the experience of the laboratorian [3].

Most cases of malaria occur in people living in or traveling to a tropical country or region. Africa carries a disproportionately high share of the global malaria burden. In 2021 the Region was home to about 95% of all malaria cases and 96% of deaths. Children under 5 years of age accounted for about 80% of all malaria deaths in the Region. Four African countries accounted for just over half of all malaria deaths worldwide in 2021: Nigeria (31.3%), the Democratic Republic of the Congo (12.6%), United Republic of Tanzania (4.1%) and Niger (3.9%) [4].

Malaria is preventable and curable. It can often be prevented by the use of antimalarial drugs and application of protective measures against mosquito bites. Travelers from different countries may receive different recommendations, reflecting differences in protocols as well as availability of medicines in different countries. However, it is important to note that you are still at risk for malaria even with the use of protection [5].

Malaria can be a severe, potentially fatal disease (especially when caused by Plasmodium falciparum), and treatment should be initiated as soon as possible. Treatment consists of antiparasitics and other supplements that can support the body's recuperating mechanism. Choice of drug regimen depends on the clinical status of the patient, the type (species) of the infecting parasite, the area where the infection was acquired and its drug-resistance status, pregnancy status, and finally history of drug allergies, or other medications taken by the patient [6]. WHO maintains a list of medicines that are used as first-line treatment in endemic countries for uncomplicated and severe malaria, as well as for prevention and treatment during pregnancy. Through its various expert groups, it regularly reviews evidence on current and new treatments to ensure that its recommendations are based on the most recent evidence. New and updated recommendations are published in the WHO guidelines for malaria. These consolidated guidelines bring together all of WHO's current recommendations for malaria they are intended as a living resource and are updated periodically as and when new evidence becomes available. WHO also supports Member States to translate these recommendations into national policies as well as to ensure their effective implementation [7].

PROBLEM STATEMENT

WHO's Global Technical Strategy for malaria has highlighted the importance of malaria surveillance as the third pillar for moving closer to malaria elimination [8]. 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 have identified errors in their data handling, modeling techniques and inclusion of critical factors that affect malaria dynamics. To tackle these limitations, we (Team Gitlab) want to include new measurement indices and bring in fresh perspective to their malaria prediction model. We will also study the possible benefits of interactions amongst 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.

RESEARCH 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 complimentary 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.

LITERATURE REVIEW

Machine Learning offers the ability to extract knowledge from data to identify relevant patterns using classification. These patterns aid in decision-making. Various studies have been conducted for predicting malaria incidence using machine learning techniques.

The analysis of the effects of climatic factors on malaria incidence in some West African countries was carried out using a statistical model that showed a negative correlation for temperature, rainfall, and malaria incidence in some areas and a positive association for other areas [9]. Another study was carried out to identify the association between climate factors and malaria incidence in Ethiopia using Pearson's correlation method. The results showed a positive relationship between rainfall and relative humidity in one part of the country, whereas the results from other regions showed an insignificant relationship between them [10]. A study published in 2022 which tried predicting malaria using deep learning models and city clusters in the state of Amazonas, Brazil, suggested that ML and DL models can be potentially low-cost decision support tool for supporting national, regional, and local malaria control strategies [11]. A data-driven malaria epidemic early warning system that can predict the 13-week case rate in a primary health facility in Burkina Faso was built using machine learning. Results of this study showed that features such as the absolute number of consultations and the variance are a good predictor of the expected daily rate of malaria cases [12]. Stephen Adebanji and his team in Osun State University built a model for predicting malaria outbreak using machine learning technique. Their research showed that Naive Bayes algorithm can be used to develop a model for predicting malaria outbreak using malaria incidence data and meteorological data [13]. Another study was carried out on building a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System. The primary results obtained in this study demonstrated the power of the proposed predicative analytics-based malaria outbreak warning system. The technology can provide alternative solutions by allowing for early warning mechanisms to monitor the spread of disease and advance management of treatment facilities to ensure a more timely health services that can save lives. The availability of any predictive model will not only help healthcare services but also to avoid or reduce the large-scale spread of diseases. The further development of the system will incorporate automatic data gathering from a variety of sources [14]. Brown, B.J., Manescu, P., Przybylski, A.A. et al. worked on building a data-driven malaria prevalence prediction tool for use in large densely populated urban holoendemic sub-Saharan West Africa with low error. The Region-specific Elastic Net based Malaria Prediction System (REMPS) they built showed good generalization performance, both in magnitude and direction of the prediction, when tasked to predict monthly prevalence of previously unseen data from years 2015, 2016 and 2017.

Many studies have been done on using machine learning techniques to improve malaria prediction, considering various influencing factors including climate variables, population data and standard malaria preventive measures. These studies all aim at developing a cutting-edge solution that can enhance decision-making in malaria control in Africa. This paper presents a research on improving the performance and interpretability of one of these models, built by Team Flask of the Hamoye Winter cohort. It outlines our procedures for analyzing the state of malaria in African countries and determining new features to engineer in the existing model. It considers the temporal and spatial patterns of malaria incidence and other malaria metrics. It classifies and maps out malaria incidence in African countries into high, medium and low incidence regions based on reported incidents in a ten-year period and goes ahead to propose differing approaches towards reducing malaria incidence rates in these regions based on model predictions. Also, it models a counterfactual scenario of malaria incidence rates in the absence of preventive measures and considers the possibility of leveraging on combinatorial effect of various malaria prevention strategies to reduce malaria incidence in Africa. Relevant datasets that contribute to the specific numerators and denominators of malaria metrics were obtained from world bank data reports. We extract the data on nationality, malaria incidence, confirmed cases, use of preventive measures and geometry.

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 Wrangling / Preprocessing: During the data preprocessing phase, extensive cleaning, merging, and wrangling were performed to ensure data quality and consistency. The original dataset used by Team Flask had a substantial number of missing values in relevant columns. To address this issue, a systematic approach was employed. The newly obtained datasets, relevant to each column with missing data, were utilized to impute these missing values. This process involved a comprehensive data wrangling strategy and the application of the `combine_first` method in the Pandas library. This approach was chosen to ensure a more accurate and precise representation of the data, rather than resorting to discarding columns or replacing missing values with zeros.

Furthermore, additional variables, such as population figures for both urban and rural areas, total population, malaria-related deaths, and estimated numbers of confirmed malaria cases, were incorporated into the dataset. This integration involved melting the new datasets to align with the original dataset structure and then utilizing an inner join operation to merge the relevant data.

To enhance the readability and usability of the preprocessed data, a reshaping and reordering process was executed. This step ensured that the dataset was structured in a logical and coherent manner, facilitating easier analysis and interpretation.

DATA ANALYSIS

The data analysis and feature engineering stage of this work represent a substantial portion of the work, given the ambitious objectives. The addition of new datasets also necessitated further analysis. Below are the stages and processes of the analysis:

Spatio-temporal Analysis:

Upon a thorough examination of the cleaned data, it became evident that the dataset contained both spatial and temporal information. Some preprocessing was undertaken to leverage these information. Specifically, the column containing spatial information, denoted as ‘latitude’, ‘longitude’ and `'geometry'`. The geometry point format was transformed into a 'polygon' type. This conversion was achieved by integrating a shapefile of Africa with the cleaned dataset. This fusion of spatial and temporal information fundamentally influenced the analytical approach. Broadly, spatial and temporal analyses were conducted at two key levels

Population Level:

In this stage of analysis, the focus was on malaria metrics related to the population which forms the denominator for most of engineered features. Several engineered features were created to provide a deeper understanding of malaria's impact. These features encompassed *standardized incidence rates* **-** calculated to account for population variations; mortality rates - an important metric reflecting malaria-related deaths; *prevalence rates* **-** indicating the proportion of the population with malaria; *case fatality rates* **-** expressing the ratio of deaths to the number of confirmed malaria cases; *total malaria cases* **-** an aggregate measure of the malaria burden.

Incidence Rates Analysis:

In the pursuit of understanding malaria incidence rates across African countries, a temporal trend analysis was conducted, and the findings were visually represented using a small multiples grid. This approach efficiently captured the incidence rates for all African countries over the specified time frame. The insights derived from this visualization led to the classification of these countries into three distinct groups based on their incidence rates: high, medium, and low. This classification was accomplished by establishing threshold values using percentiles and employing a logical control mechanism, often facilitated by a for loop, to categorize the countries.

Furthermore, a classification approach was employed to identify countries with common high, medium, or low incidence rates across each year. The classification of countries into high, medium, and low incidence rate groups based on their temporal trends prompted an investigation into regional comparisons. Specifically, the goal was to understand how these incidence rates compared on average across all countries in Africa over the years. The findings from this analysis consistently aligned with the countries categorized as high and medium incidence countries. This suggests that the regions exhibiting high and medium incidence rates contribute significantly to the overall malaria burden in Africa.

In addition to the regional analysis, the impact of subpopulations, particularly rural and urban populations on the incidence was explored.

The addition of the population dataset enabled us to perform more robust feature engineering, especially in terms of adjusting the incidence rates to account for changes in population. Previously, the project had considered incidence rates calculated from an unknown population, but with the new population dataset, it became possible to calculate 'Standardized Incidence Rates.' These standardized rates were computed using the formula:

Standardized Incidence Rates = $\frac{Number\ of\ New\ Cases}{Total\ Population\ at\ Risk} \times 1000$

The same analytical techniques that were applied earlier to the raw incidence rates were used for the standardized incidence rates.

Geospatial analysis was conducted to visualize and map countries with high, medium, and low incidence rates over the years. This spatial representation provides an intuitive way to understand the geographic distribution of malaria incidence rates.

Mortality Rate Engineering:

In addition to incidence rates, another important malaria metric was engineered based on population data: the mortality rate. The mortality rate was calculated using the formula:

Mortality Rate = $\frac{Number\ of\ Malaria\ Deaths}{Total\ Population\ at\ Risk} \times 1000$

Similar to the methods performed on incidence rates, we classified countries into low, medium, and high mortality rate categories. This classification was instrumental in identifying countries with both high mortality and high incidence rates, countries with medium mortality and high incidence rates, and countries with low mortality and high incidence rates. Temporal trend, regional comparison and the effect of subpopulation of the mortality rates were also conducted using the approach stated earlier.

Case Fatality Rate (%):

The project's population-based analysis included the engineering of another significant feature, the 'Case Fatality Rate.' This rate is a crucial metric as it measures the proportion of individuals who die from a specific disease among all individuals diagnosed with that disease during a certain time period. It serves as an indicator of disease severity and is often utilized for prognosis, with higher rates signifying relatively poor outcomes. Additionally, the Case Fatality Rate is useful for evaluating the impact of new treatments, with rates expected to decrease as treatment options improve. Like other malaria metrics, it is population-dependent and varies across populations and over time. The formula used to calculate the Case Fatality Rate is: $\frac{Number\ of\ Malaria\ Deaths}{Malaria\ Confirmed\ Cases} \times 100$.

For this particular analysis, a dataset containing the necessary features for calculating the Case Fatality Rate was imported, covering the years from 2010 to 2017 (distinct from the 2007-2017 period used for other malaria metrics). From here the temporal trend of case fatality rate was performed.

Given that Case Fatality Rate (CFR) is not only indicative of disease severity but can also be used to evaluate the effectiveness of new treatment, we expanded its analysis to explore the relationship between CFR and other variables. We considered variables related to preventive measures that includes the population of people using Insecticide-Treated Nets (ITNs), Intermittent Preventive Treatment in Pregnancy (IPTp), and Antimalarial drugs.

A correlation analysis was conducted to investigate the relationships between CFR and these preventive measures across the years. This analysis could help in assessing whether the utilization of preventive measures, such as ITNs, IPTp, and antimalaria drugs, has any observable impact on reducing the severity of malaria, as reflected in the CFR.

Preventive measures level

In this phase of the analysis, we delved into the impact of various preventive measures on reducing malaria incidence rates in Africa over the years. Nine features related to preventive measures, drawn from population data, were selected for investigation. A correlation analysis was initiated, with a predefined correlation threshold of 0.3 using Pearson correlation coefficients. The goal was to examine the degree of correlation among all preventive measures and identify highly correlated features among both the preventive measures themselves and the malaria metrics. The results obtained from this analysis does not indicate causation. While positive correlations suggest that certain combinations of preventive measures might be effective, further analyses and studies are needed to establish causal relationships

The information obtained from the correlation analysis was then used to estimate the total usage counts of preventive measures associated with basic drinking water services and basic sanitation services across all countries and years was estimated. The results from this estimation suggests that there may be other confounding variables beyond the scope of this project contributing to the incidence rates in some countries in Arica. It also underscores the complexity of the malaria issue and the need for multifaceted interventions, potentially including vaccination, in countries with persistently high malaria rates.

Counterfactual Scenarios of Preventive Measures

To assess the impact of preventive measures, we modeled a counterfactual scenario using a random forest regressor in the absence of preventive measures. This counterfactual scenario aimed to predict incidence rates higher than the actual incidence rates due to the absence of preventive measures. The resulting estimates predicted from the counterfactual scenario was compared with the actual incidence rates. A positive difference between the actual and counterfactual suggests that the total number of people using preventive measures in that year did not lead to a decrease in incidence rates or mortality rates, whereas a negative difference indicates that the observed incidence rate with preventive measures was lower than what would have been predicted without them.

Information obtained from here was used to determine which preventive measures had the greatest impact

Spatial analysis

We utilized various libraries, including Folium, PySAL, Libpysal, and ESDA, for spatial analysis. This analysis aimed to explore the spatial distribution and clustering patterns of malaria incidence rates and mortality rates. First,

spatial clustering analysis was performed to detect spatial patterns in malaria incidence rates and mortality rates using spatial autocorrelation techniques. Two forms of analysis were employed:

1. Global Moran's I's Statistics:

Global Moran's I's Statistics were used to determine the degree of spatial pattern in malaria cases in Africa. This statistic measures the correlation between a variable at a location and the values of the same variable at neighboring locations. It can indicate whether the data is clustered, dispersed, or random.

2. Hotspots / Cold Spots Analysis using Local Getis-Ord Statistics (Gi):*

Hotspots and cold spots analysis using the Getis-Ord Gi* statistical analysis was performed to identify specific spatial clustering patterns, such as hotspots (areas with high values) and cold spots (areas with low values), with statistical significance. This analysis considers the local mean of cases for a country and its nearest neighboring country in relation to the global mean for all countries in Africa. The 'z' score obtained indicates high or low malaria incidence clusters.

Hotspots analysis was conducted on a yearly basis as well as for the average over the years. The 'z' score's magnitude indicates the strength of the relationship, with scores closer to zero indicating the absence of clusters.

Hotspots analysis was also used to identify countries with both high incidence and high mortality rates, as shown.

Spatial regression was also conducted using the **spreg** module from the PySAL library to investigate the relationships between population, mortality rate, incidence rate, and total malaria cases

MODEL DEVELOPMENT

In pursuit of our overarching goal to enhance the performance and interpretability of our predictive model, the model development phase integrates insights derived from the preceding analysis and feature engineering stages. Feature selection was carried out to ascertain the most salient features. We harnessed the permutation feature importance technique. This technique, in conjunction with the Random Forest Regressor and XGBOOST.

Model Building

Four models were used. These include the Random Forest Regressor and XGBOOST. A Stacking Regressor and a Voting Regressor which combines the individual regressors (Random Forest and XGBOOST) were also used. The process used for model building followed the regular approach of standardizing the input variables using StandardScaler, splitting the data into training and testing 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-squared.

Evaluation of Predictive Accuracy

Ten-fold cross validation was performed on each of the models after which hyper-parameter tuning was done using the random search CV.

Model Interpretation

Model interpretability constitutes a cornerstone of our research objective. To this end, we employed pertinent techniques and tools to facilitate a deeper understanding of our predictive models. The foremost tool utilized for this purpose was LIME, which provides a localized interpretation of predictions for specific data points. This approach involves defining the feature names, initializing the LimeTabularExplainer with feature names, choosing individual data point for explanation, defining a base classification threshold, classifying the predicted incidence rates as high, low or medium using control statements, and finally showing the explanation in notebook

Model Deployment

The best performing models was deployed using the Streamlit app.

RESULTS

Malaria metrics results

Incidence rates – . Figure 1.1 to Figure 1.3 shows various graphical representations of malaria incidence and prevalence in Africa. Figure 1.1 is a bar graph of total malaria incidence in African countries for the ten-year period under review. Figure 1.2 is a bar graph of malaria incidence rates. Figure 1.3 show a line chart for displaying trends and change over time of malaria prevalence.

From the classification approach to classify countries into high, low, and medium incidence countries Figure 1.3 shows the cluster map that clustered countries that shared similar incidence rates, revealing a hierarchical clustering pattern. This facilitated the identification of countries with consistently high incidence rates over the years, including Benin, Burkina Faso, Nigeria, Mozambique, Liberia, and the Central African Republic.

Considering the effect of the subpopulations on the incidence rates, it can be seen in Figure 1.5 that malaria incidence is higher in the urban areas than in the rural areas.

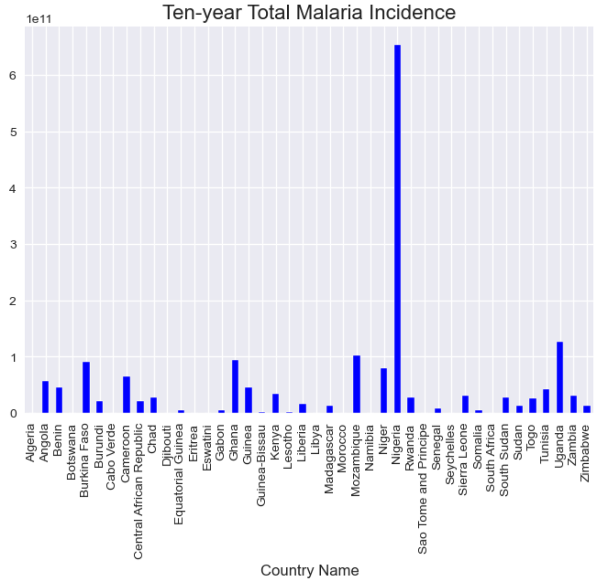


Figure 1.1: Distribution of the total malaria incidence

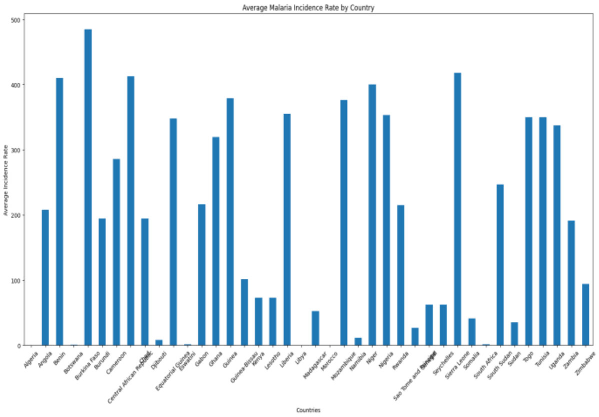


Figure 1.2: Bar Graph of Malaria Incidence per 1000 population

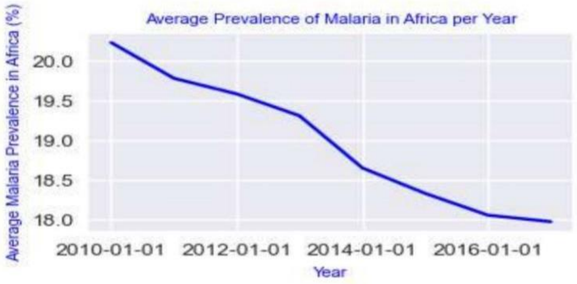


Figure 1.3 Line Graph showing annual average of malaria prevalence in Africa between 2007 and 2017

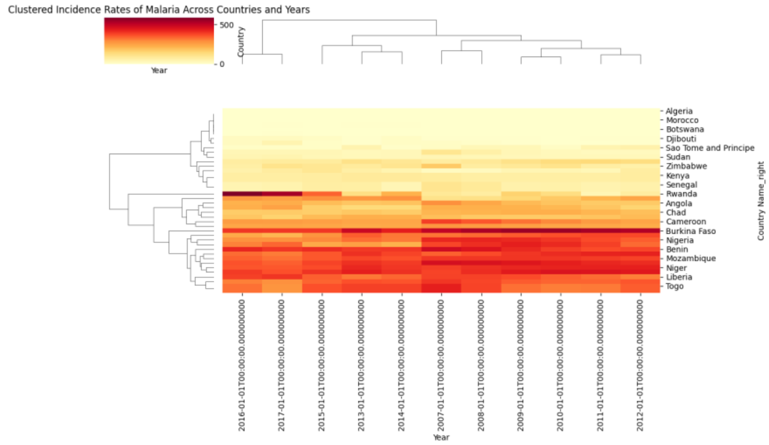


Figure 1.4: Clustermap of showing the incidence rates of countries based on similarities

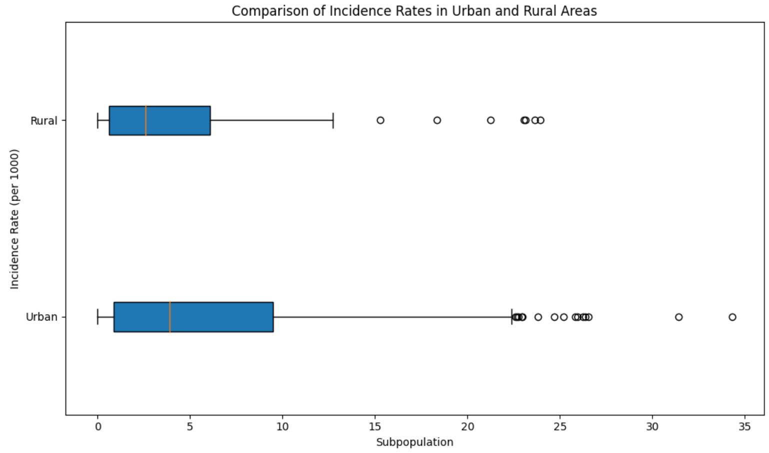


Figure 1.5: A Box Plot showing the distribution of incidence rates in urban and rural areas.

Mortality rates – Over the years, a decrease in mortality rates was observed, indicating progress in the fight against malaria as seen in Figure 2.1. Just like the incidence rates, the classification approach employed classified some countries that intersect between mortality rates and incidence depending on their degree.

The result is shown below

```
Countries with Both High Mortality and High Incidence Rates: {'Liberia', 'Burkina Faso', 'Mozambique'}
Countries with Both Medium Mortality and Medium Incidence Rates: {'Senegal', 'Burundi', 'Angola', 'Zambia', 'Guinea-Bissau', 'Madagascar', 'Kenya', 'Cameroon', 'Sudan', 'South Sudan', 'Gabon', 'Togo', 'Equatorial Guinea', 'Ghana', 'Zimbabwe', 'Namibia'}
Countries with Both Low Mortality and Low Incidence Rates: {'Libya', 'Morocco', 'Botswana', 'Eswatini', 'Sao Tome and Principe', 'Algeria', 'South Africa'}
```

The effect of the mortality rates on the subpopulation is shown in Figure 2.2

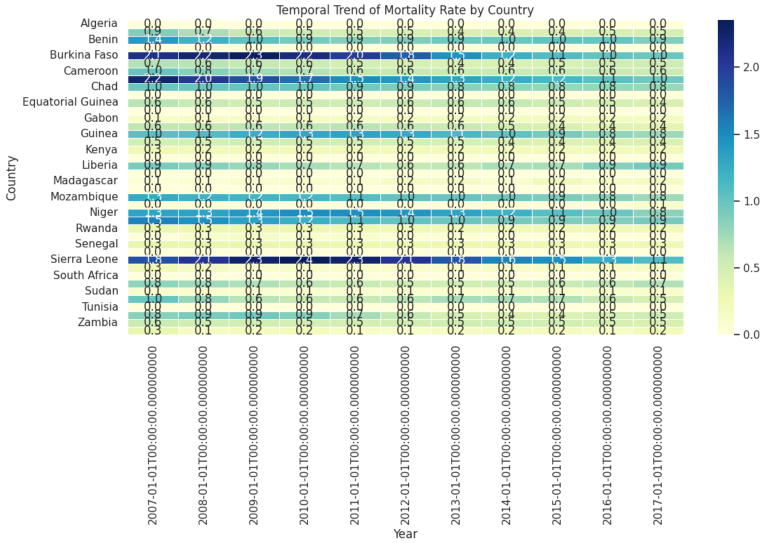


Figure 2.1: A temporal trend of mortality rates by country and year

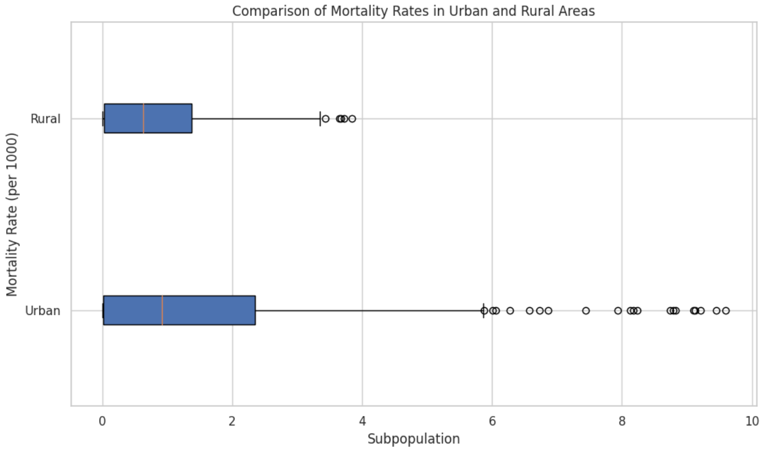


Figure 2.2: A boxplot showing the effect of mortality rates in the subpopulations

Case Fatality Rate (CFR): Figure 3.1 and 3.2 shows the temporal trend of CFR over years. By this, the countries that have the highest CFR over the year is shown in Figure 3.3 From 3.1 and 3.2, it can be observed that the highest CFR occur in the year 2010 and 2014. This observation led to the result in Figure 3.4 and Figure 3.5 to know the countries that have the highest CFR in those years. Countries with the Highest CFR in 2010 include Cameroon, Central African Republic, and Guinea had the highest Case Fatality Rates. Also in 2014, Cameroon still had the highest Case Fatality Rate, and Central African Republic and Guinea, while reduced, remained among the top 10 countries with high Case Fatality Rates.

Furthermore, correlation analysis done between CFR, incidence rates and the percentage of the population using one preventive measures is also depicted in Figure 3.6

Counterfactual Scenarios of Preventive Measures

Figure 4.1 shows a subplot of the actual incidence rates with preventive measures and the counterfactual predicted incidence rates without preventive measures using the Random Forest Regressor. It is observed generally that some percentage of people that used certain preventive measures in year 2007, 2010 and 2011 proved effective in reducing the malaria incidence rates. The preventive measures that contributed the most to this counterfactual scenario are shown in Figure 4.1 based on their feature importance. The result obtained from the geospatial analysis using the incidence difference to map regions where preventive measures were effective and those where they did not is shown in Figure 4.3

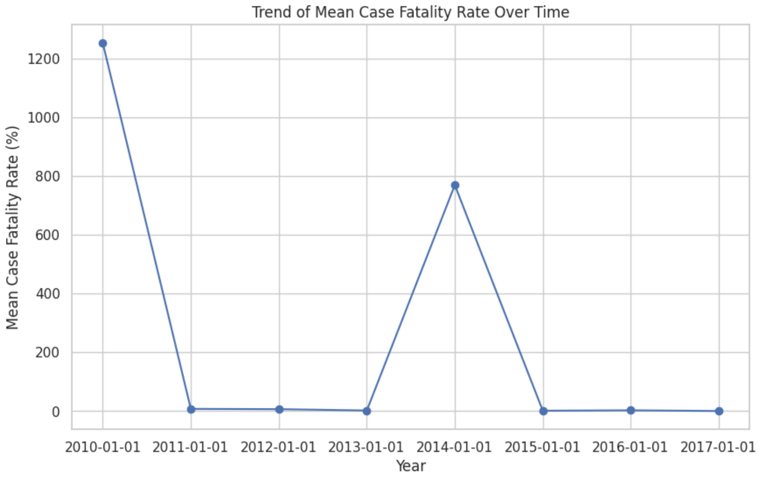


Figure 3.1: A temporal trend of Case Fatality Rate

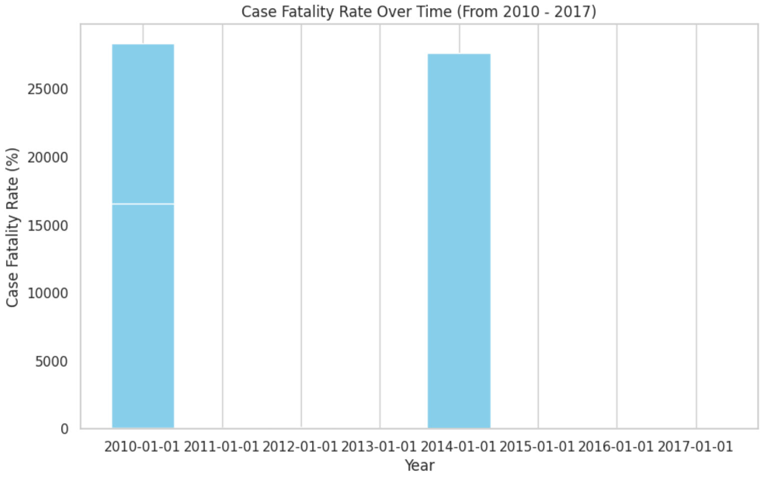


Figure 3.2: Distribution of CFR over time

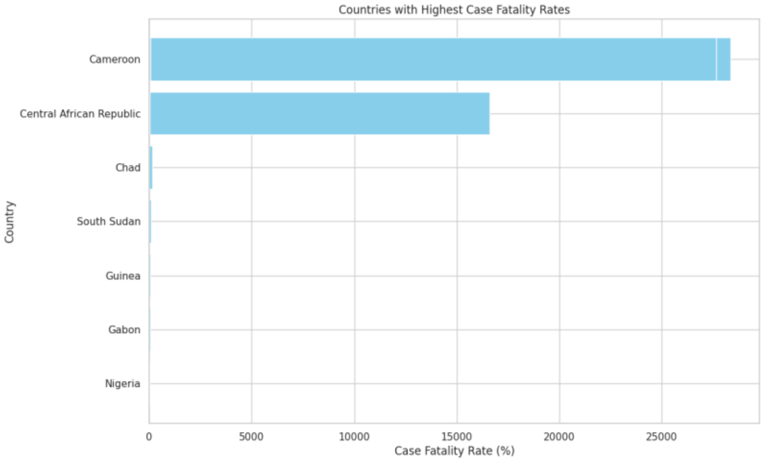


Figure 3.3: Distribution of countries with the highest CFR across the years

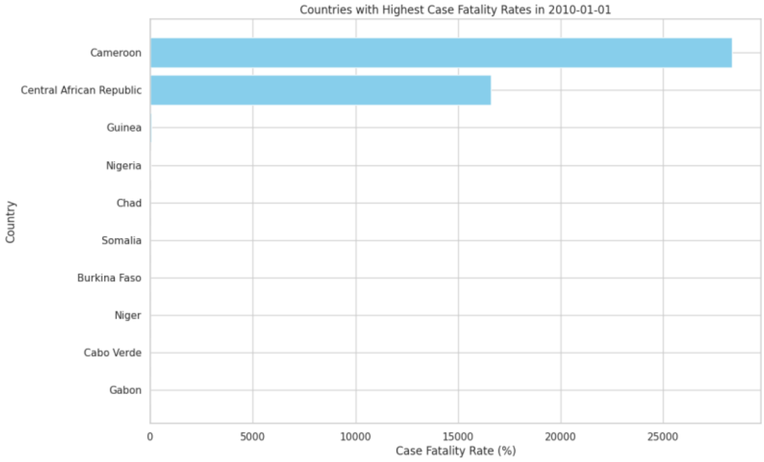


Figure 3.4: Distribution of countries with the highest CFR in 2010.

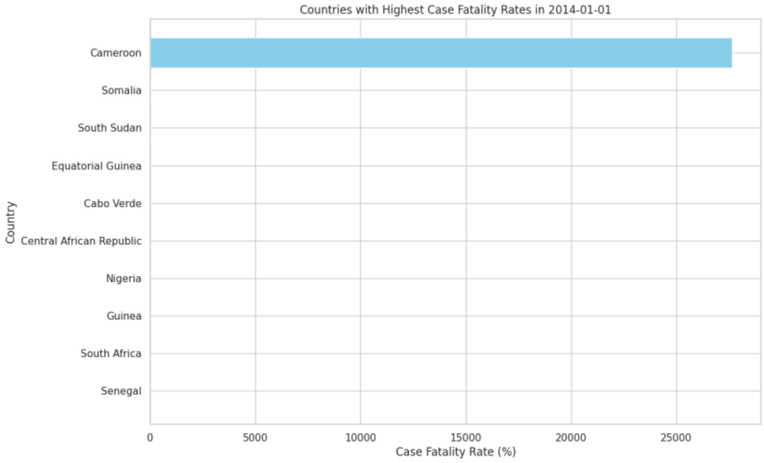


Figure 3.5: Distribution of countries with the highest CFR in 2014.

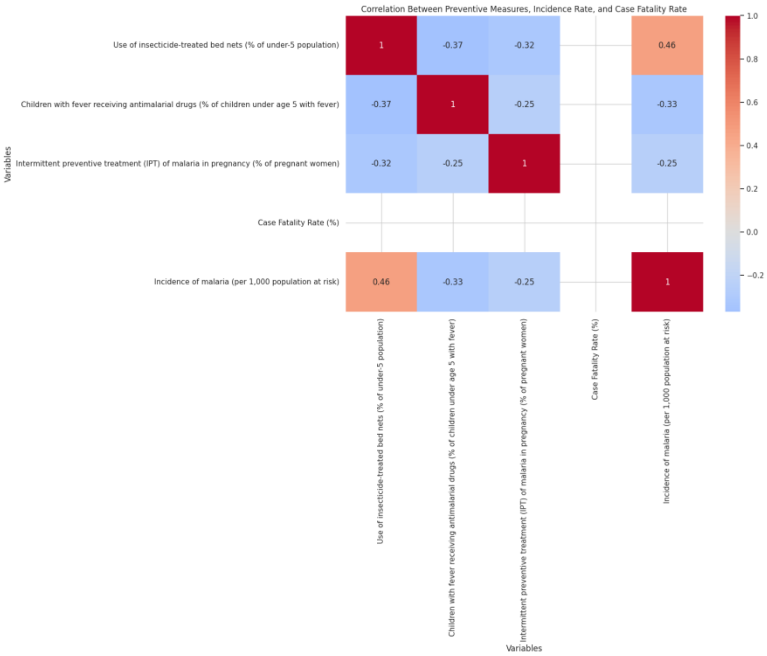


Figure 3.6: A correlation matrix between CFR, incidence rates and % of population using certain preventive measures.

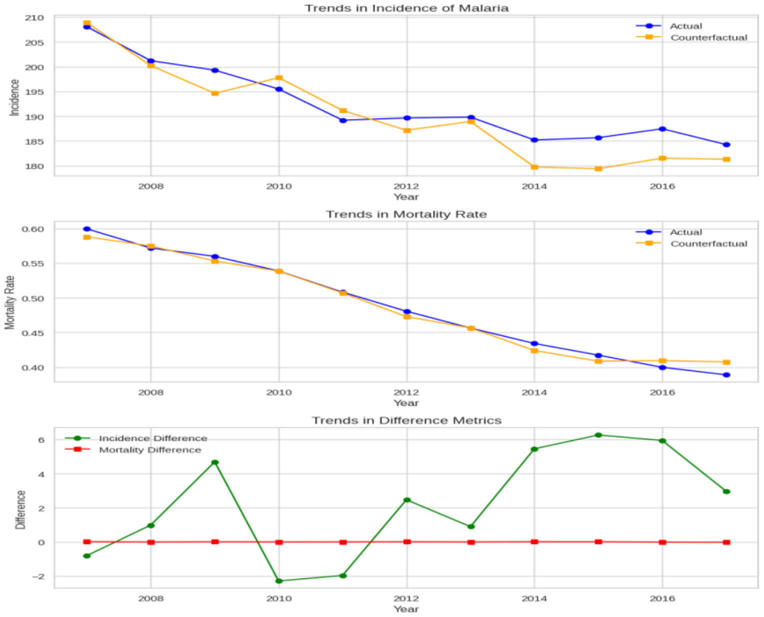


Figure 4.1: Counterfactual scenarios of malaria incidence rate

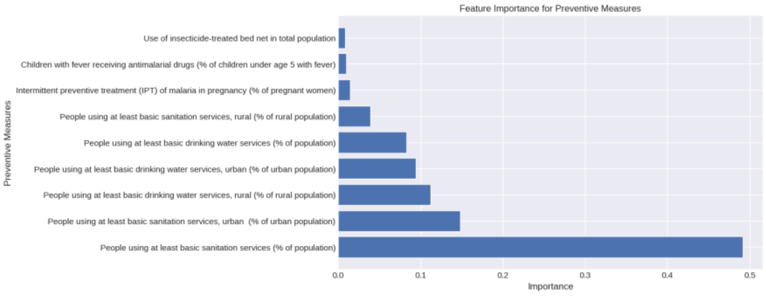


Figure 4.2 : Feature importance of preventive measures

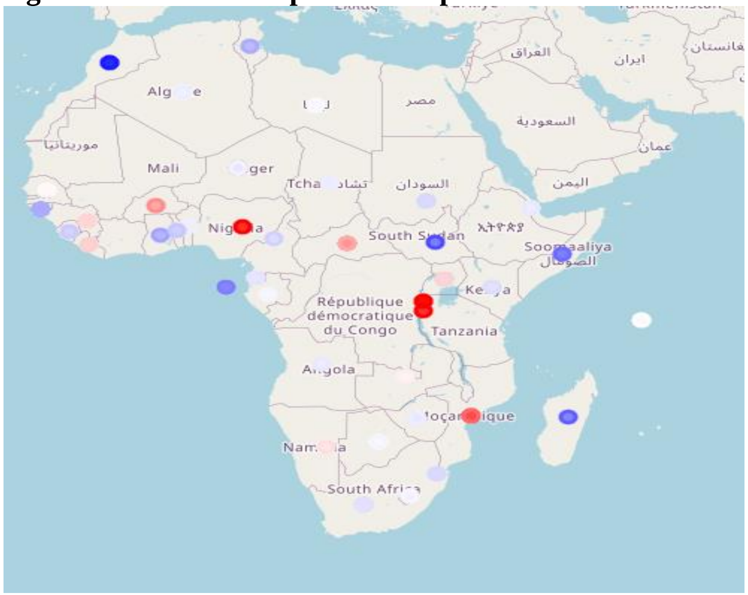


Figure 4.3: Geospatial analysis of hotspots and coldspots areas

Spatial autocorrelation: The result from the spatial analysis using the spatial autocorrelation techniques observed a negative autocorrelation of -0.03 from the Global Moran's I's statistics. This is shown in Figure 5.1. This is statistically significant as p-value < 0.05. The result obtained from the second part of the analysis, that is, Getis-Ord Gi* statistics is shown in Figure 5.2. This reveals the present of hotspot and cold areas based on incidence rates. Hotspot analysis using the Getis-Ord Gi* based on high incidence and high mortality rates countries is also shown in Figure 5.3. Hotspots areas are indicated as red while cold spots areas are indicated as blue depending on the Gi*

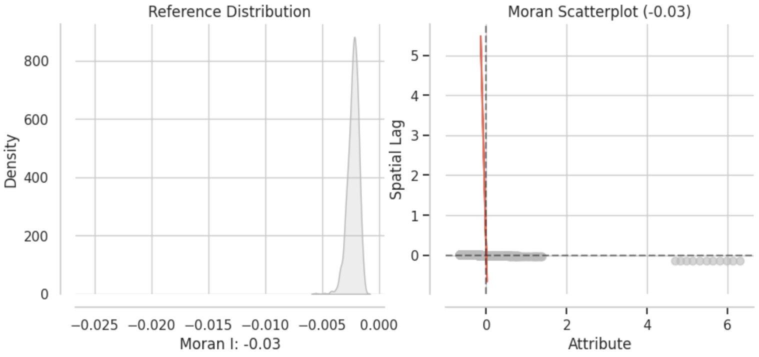


Figure 5.1: Moran's plot

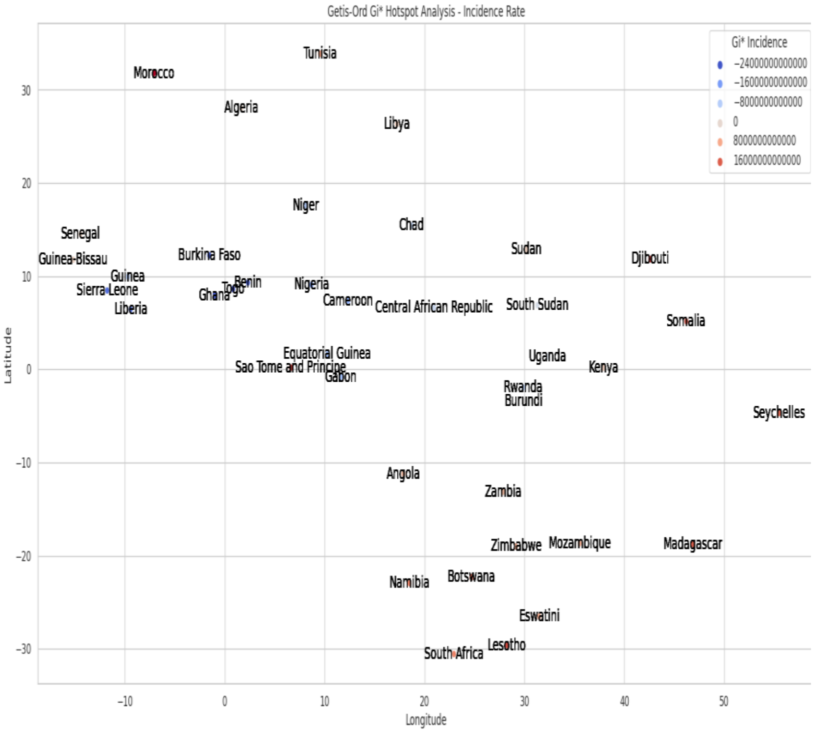


Figure 5.2: Hotspot analysis based on incidence rates on average

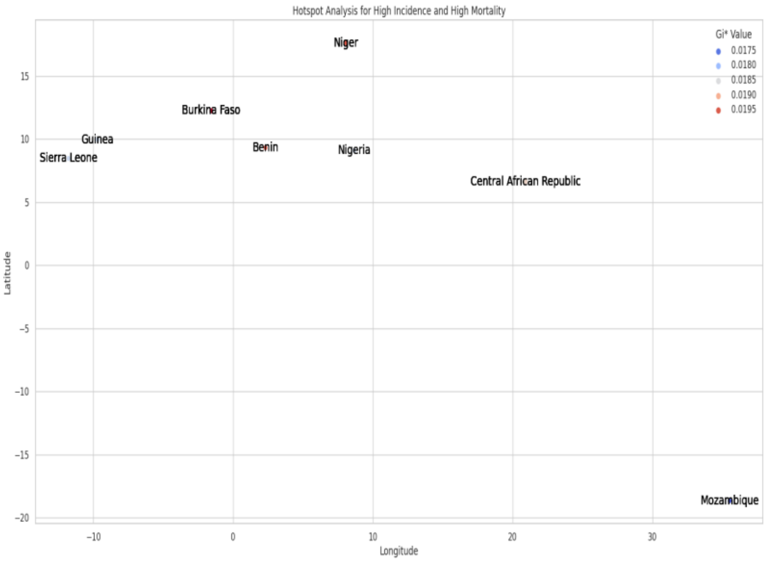


Figure 5.3: Hospot analysis for high incidence and high mortality rate countries

Model results and predictive performance

To estimate the relationship between the predictor variables and malaria incidence and remove multicollinearity, feature selection was done and the results using permutation feature importance using RMSE as error metrics on Random Forest Regressor and XGBOOST is shown in Figure 6.1. This approach guided in the selection of features (other engineered features were also added). Once accounting for other predictors, the percentage of people using preventive measures such as ITNs, IPTp and antimalarial drugs were not captured by the models during permutation feature importance techniques. These predictors were dropped.

Predictive performance in the training and test sets, cross validation on training set and hyperparameter tuning is presented in table 1. The baseline predictive performance on the training sets were excellent for all the models used with R2 score within the range of 0.99. As expected, the predictive performance on the test sets dropped a bit with R2 score ranging from 0.93 to 0.99 across all the models used. Model diagnostics (as presented in Table 2) in form of cross validation and hyper-parameter tuning was very satisfactory. For the cross validation on the baseline models, predictive performance on the training set still remained high as the R2 score was within the range of 0.91 – 0.95 in all models and hyperparameter tuning was very satisfactory as well, The predictive performance, as expected, increased when evaluated on the test sets as R2 score ranged from 0.97 to 0.99. The Random Forest has the highest performance and this was chosen as our best model. The performance of individual models is presented in Figure 6.3

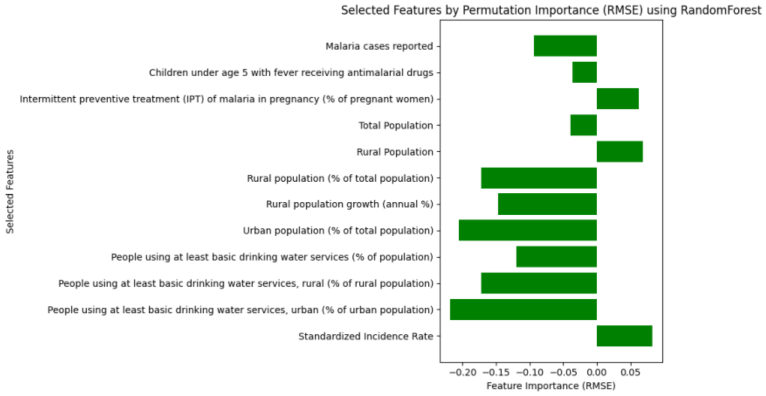


Figure 6.1: Permutation feature importance for feature selection using Random Forest

Model interpretation

The result obtained from LIME is presented in Figure 7.1 and 7.2. As LIME gives a local interpretation of a specific data point, we checked for two data point to get an explanation of how the prediction was done. The data points chosen include 10, 50, 65 which were predicted as low, medium and high respectively. This is shown in Figure 7.1, 7.2, 7.3.

Model Deployment

The model was deployed on streamlit

Table 1: Baseline Model Performance on both train and test set

	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

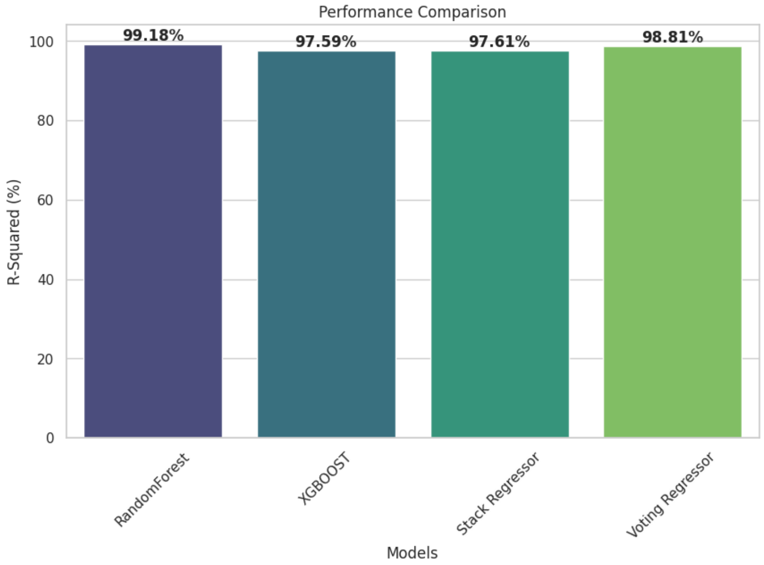
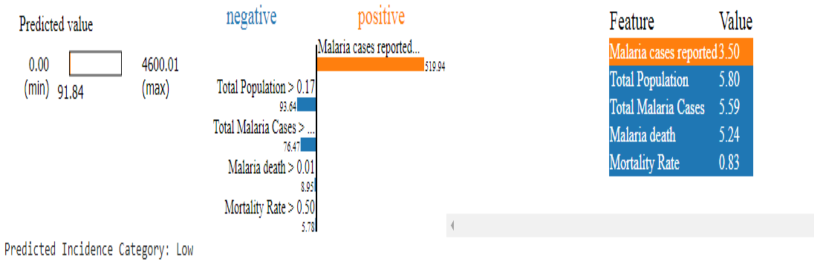


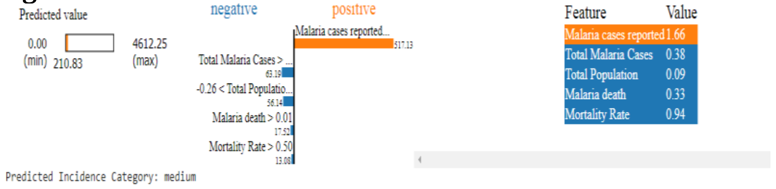
Figure 6.3: Performance comparison among models



Feature	Value
Malaria cases reported	3.50
Total Population	5.80
Total Malaria Cases	5.59
Malaria death	5.24
Mortality Rate	0.83

Predicted Incidence Category: Low

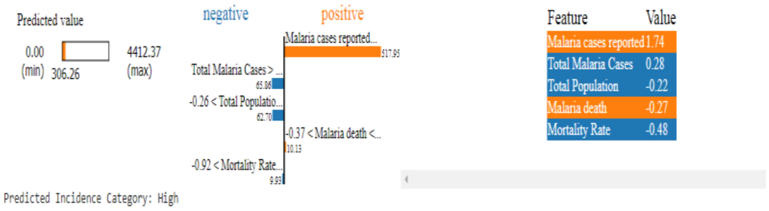
Figure 7.1: LIME explanation at data point 0 predicted as high



Feature	Value
Malaria cases reported	1.66
Total Malaria Cases	0.38
Total Population	0.09
Malaria death	0.33
Mortality Rate	0.94

Predicted Incidence Category: medium

Figure 7.2: LIME explanation at data point 50 predicted as medium.


Predicted value
0.00
(min)
306.26
4412.37
(max)
negative
positive
Malaria cases reported...
517.95
Total Malaria Cases > ...
65.86
-0.26 < Total Populatio...
62.70
-0.37 < Malaria death <...
10.13
-0.92 < Mortality Rate...
9.93

Feature	Value
Malaria cases reported	1.74
Total Malaria Cases	0.28
Total Population	-0.22
Malaria death	-0.27
Mortality Rate	-0.48

Predicted Incidence Category: High

Figure 7.3: LIME explanation at data point 65 predicted as high

DISCUSSION

The pursuit of constructing a predictive model for malaria incidence rates, with the overarching goal of advancing our understanding of control strategies, has culminated in a profound exploration of data, meticulous analysis, and the development of predictive models. In this discussion, we delve into the nuanced inferences and multifaceted contributions that have emerged throughout this research journey.

Our journey commenced with data preparation and wrangling, where the amalgamation of diverse data sources from WHO, UNICEF and the World bank posed unique challenges. Addressing missing values with judicious imputation strategies was vital, highlighting the paramount importance of data completeness in constructing reliable predictive models. The temporal scope of the dataset, spanning 2007 to 2017, enabled the analysis of long-term trends and dynamic control strategies.

The data analysis and feature engineering phase unveiled several pivotal insights. Notably, a declining trend in both incidence and mortality rates over the years signifies encouraging progress in malaria control, although disparities across countries underscore the need for tailored strategies. Countries with persistently high incidence rates, including Benin, Burkina Faso, Nigeria, Mozambique, Liberia and the

Central African Republic, demand targeted interventions. Higher incidence rates in urban areas suggest urbanization's role in malaria transmission dynamics.

Features engineering enriched the dataset with population-based metrics, introducing Standardized Incidence Rates, Mortality Rates, and Case Fatality Rates. These metrics empower stake holders with additional tools for strategic interventions.

The Case Fatality (CFR) representing the proportion of malaria-related deaths among diagnosed cases, emerged as a critical metric. Temporal analysis revealed peak CFRs in 2010 and 2014, signifying temporal variations in disease severity. Also, the heightened CFR in these years demand a closer examination to uncover the factors contributing to this temporal variation. Further investigations into the countries with the highest CFR in 2010 and 2014 yielded additional revelations. Cameroon, Central African Republic and Guinea were identified as countries with the highest CFR in 2010. In 2014, Cameroon retained its unenviable position, while Central African Republic and Guinea exhibited a notable reduction in CFR, albeit remaining within the top 10 countries with high case fatality risk. Furtthermore, spatial analysis unlocked the spatial patterns inherent in malaria incidence and mortality rates. A pivotal inference here is the presence of negative spatial autocorrelation in incidence rates. This observation challenges the notion of random disease dispersion, hinting at underlying factors that influence the spatial distribution of cases. The non-random pattern, where high incidence countries are surrounded by low incidence neighbours , could be an avenue for future research. Also, the identification of hotspots and cold spots through clustering analysis serves as a geographical compass for intervention strategies. These delineation provides valuable guidance on where resources should be channeled and where targeted measures can have the most substantial impact. The identification of countries simultaneously classified as hotspots for both high incidence and high mortality underscores the gravity of the situation in these regions, necessitating multifaceted interventions.

Model performance in predicting incidence rates was excellent with the best performing model having R^2 score of 99%. This has a greater performance when compared to the Random Forest used by Team Flask having R^2 score of 84% Likewise, our commitment to model interpretability embodied in the application of LIME, facilitates a deeper understanding of our predictive model. The inferences drawn from these interpretability efforts are pivotal. They bridge the gap between the enigmatic nature of machine learning models and actionable insights. By classifying predictions into meaningful categories and elucidating the contributing factors for specific data points, we empower stakeholders with the ability to make informed decisions grounded in model outputs.

Conclusion

In all, the analysis and approaches undertaken in this study contributes to our comprehension of malaria incidence dynamics, particularly in the context of Africa. The identification of hotspots regions with consistently high mortality and incidence rates informs precision control strategies. This research guides the allocation of resources to areas that need them most, optimizing the impact of control efforts. Also, by achieving a high R^2 score, the model represents a significant scientific advancement has it has reinforced the use of machine learning approached in critical areas such as healthcare. However, limitations exist, including data quality issues, ever changing data distribution / shifts.

Acknowledgments: We thank Hamoye for giving us a platform to gain research experience as part of the internship. We also extend our gratitude to WHO, UNICEF and the World Bank for providing valuable data sources for this research.

Contributors: All authors conceived of the study with input CCM led data collection, supported by GOD and with input from KK. BTA created the repo with CCM, GOD & OV contributors. GOD led the analysis with inputs from CCM, MC, MA, OGI, OO, AA with additional inputs from KK.GOD led the model building with inputs from OV. AA led the model deployment with support from GOD. CCM drafted the manuscript with support from GOD, KK. Presentation slides led by MA and OO with support from GOD. *All authors read and approved the final manuscript*

References

[1]. https://www.cdc.gov/malaria/about/faqs.html#:~:text=How%20is%20malaria%20transmitted%3F,taken%20from%20an%20infected%20person

[2]. https://www.health.ny.gov/diseases/communicable/malaria/fact_sheet.htm

[3]. https://www.cdc.gov/malaria/diagnosis_treatment/diagnosis.html#:~:text=Malaria%20parasites%20can%20be%20identified,the%20parasites%20a%20distinctive%20appearance.

[4]. https://www.who.int/news-room/fact-sheets/detail/malaria#:~:text=The%20WHO%20African%20Region%20continues,malaria%20deaths%20in%20the%20Region

[5]. https://stanfordhealthcare.org/medical-conditions/primary-care/malaria/treatments/prevention.html

[6]. https://www.cdc.gov/malaria/diagnosis_treatment/treatment.html

[7]. https://www.who.int/activities/treating-malaria

[8]. **World Health Organization, Geneva** (2015), WHO Global Technical Strategy for Malaria 20162030, https://scholar.google.com/scholar_lookup?hl=en&publication_year=2015&author=WHO&title=Global+Technical+Strategy+for+Malaria+2016%E2%80%932030

[9]. **A. Arab, M.C. Jackson and C. Kongoli** (2014), "Modelling the effects of weather and climate on malaria distributions in West Africa", Malar J, Vol. 13, pg. 126 135.

[10]. **L. Sena, W. Deressa and A. Ali** (2015), "Correlation of climate variability and malaria: a retrospective comparative study, southwest Ethiopia", Ethiop. J. Health Sci., Vol. 25 (2), pg. 129-138.

[11]. **Barboza MFX, Monteiro KHC, Rodrigues IR, Santos GL, Monteiro WM, Figueira EAG, Sampaio VS, Lynn T and Endo PT** (2022), Prediction of malaria using deep learning models: A case study on city clusters in the state of Amazonas, Brazil, from 2003 to 2018, Rev Soc Bras Med Trop. doi: 10.1590/0037-8682-0420-2021, PMID: 35946631; PMCID: PMC9344950.

[12]. **Harvey D, Valkenburg W and Amara A** (2021), "Predicting malaria epidemics in Burkina Faso with machine learning, PLoS ONE 16(6): e0253302, doi:10.1371/journal.pone.0253302.

[13]. **Kazeem Ibrahim and Adebanji Stephen** (2021), "A Model For Predicting Malaria Outbreak Using Machine Learning Technique, Scientific Annals of Computer Science, Vol.19 https://www.researchgate.net/publication/356439342_A_MODEL_FOR_PREDICTING_MALARIA_OUTBREAK_USING_MACHINE_LEARNING_TECHNIQUE

[14]. **Modu et al.** (2017), "Towards a Predictive Analytics-Based Intelligent Malaria Outbreak Warning System", journal of applied sciences, ResearchGate Publication, Vol. 7(8):836, https://www.researchgate.net/publication/319144556_Towards_a_Predictive_Analytics-Based_Intelligent_Malaria_Outbreak_Warning_System

[15]. **Brown, B.J., Manescu, P., Przybylski, A.A. et al.** (2020), Data-driven malaria prevalence prediction in large densely populated urban holoendemic sub-Saharan West Africa, Sci Rep 10, 15918, https://doi.org/10.1038/s41598-020-72575-6

[16]. **Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N.** (2019), "Data preprocessing in predictive data mining, In Knowledge Engineering Review, Vol. 34, Issue January, https://doi.org/10.1017/S026988891800036X