HAMOYE PREMIERE PROJECT

23 MAY, 2024

FORECASTING ANTIMALARIAL DRUG NEEDS

Team: Jenkins

Our Team

Presenter

Co-Presenter

Project Lead/ Co- Presenter - Arthur Uzoma

Assistant Project Lead - Nishant Katiyar

Query Analyst 1 - Olaolu Adeniyi

Query Analyst 2 - Mohammed Sajid

Query Analyst 3/ Backup Presenter - Arnab Das

Query Analyst 4 - Toluwalope Emmanuel

Presenter/ Query Analyst 5 - Somi Fredrick

Query Analyst 6 - Folaranmi Olaniyi

Problem Statement

The stark reality of malaria's impact on child mortality necessitates robust predictions of antimalarial drug requirements leading to preventable deaths.
However, several challenges hinder the accurate prediction of antimalarial drug requirements, exacerbating treatment gaps.

- Data Challenges: Accurate forecasting requires reliable data on past malaria cases, drug consumption patterns, and environmental factors influencing transmission. Data collection in resource-limited settings can be patchy and unreliable, impacting model accuracy.

- Model Complexity: Developing robust forecasting models necessitates expertise in data analysis, modeling techniques, and malaria epidemiology. This can be a significant hurdle for low-resource healthcare systems.

- Implementation Constraints: Implementing forecasting models requires robust infrastructure and trained personnel to interpret and utilize the data effectively. This can be a challenge in regions with limited healthcare infrastructure.

Existing Solutions

1

The World Health Organization (WHO) reports on the increased mortality risk associated with stockouts of ACTs, emphasizing malaria's impact on child mortality in Africa Additionally, the CDC highlights the progression to severe malaria and the potential spread of drug resistance due to inadequate access to ACTs.

2

The CDC's research underscores the threat of drug resistance, particularly in Plasmodium falciparum and P. vivax species. WHO's Strategy to respond to antimalarial drug resistance in Africa outlines measures to mitigate the emergence and spread of resistance, emphasizing the importance of timely detection and response.

3

WHO emphasizes the importance of regular monitoring of drug efficacy and resistance to inform treatment policies and combat the spread of resistance. Two dashboards developed by WHO facilitate the collection of information on drug efficacy and molecular markers of drug resistance, supporting global surveillance efforts.

4

Medard Edmund Mswahili, Gati Lother Martin, Jiyoung Woo, Guang J. Choi, and Young-Seob Jeong conducted a project titled "Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum," aiming to predict the activity of antimalarial drugs by utilizing chemical features of compounds.
They employed binary classification based on a dataset of anti-malaria activity against Plasmodium falciparum, generating feature vectors using PaDEL- Descriptor software from simplified molecular-input line-entry system (SMILES) strings of verified experimental anti-malaria drug compounds.

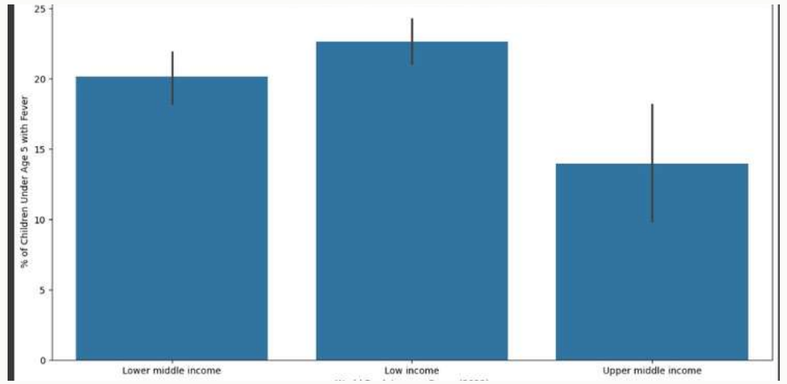
Data Description

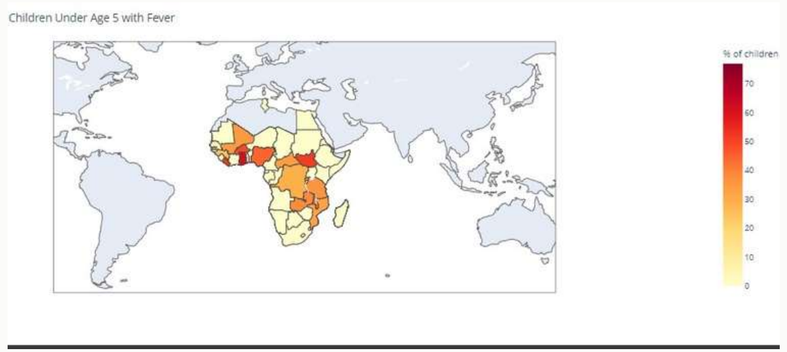
Data Sources: This dataset comprises various demographic and socio-economic indicators collected from countries and areas, classified by multiple regional and income categorizations to facilitate analysis by organizations such as UNICEF and the World Bank- World Development Indicators.

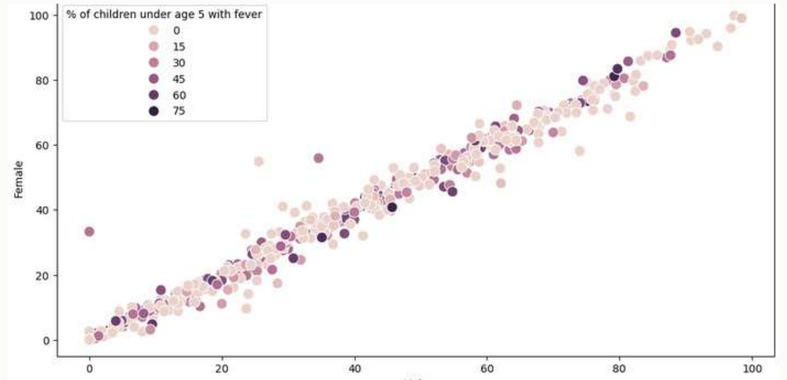
Data Features: The data includes specifics about national-level indicators, gender-based statistics, urban-rural distinctions, wealth index quintiles, and maternal education levels, along with the sources of the data.

Data Preparation: To prepare the data for analysis, we merged multiple datasets, eliminating duplicates and empty entries, removing irrelevant columns, adjusting data types for consistency, and separating numerical values from categorical ones.

Data Visualization: we used Scatter Plot, Bar Plot, Heat Map and Choropleth map to explore trends and relationships between the different features of our dataset.







Our Approach

01

Libraries and Module Importation:
We begin by importing the necessary libraries and modules to facilitate data analysis and modeling.

02

Data Preprocessing:
This stage includes cleaning and preparing the data to ensure it is suitable for modeling, addressing issues such as missing values and inconsistencies.

03

Descriptive Analysis:
This step involves summarizing and interpreting the data to understand its key characteristics and gain insights into the underlying patterns and trends.

04

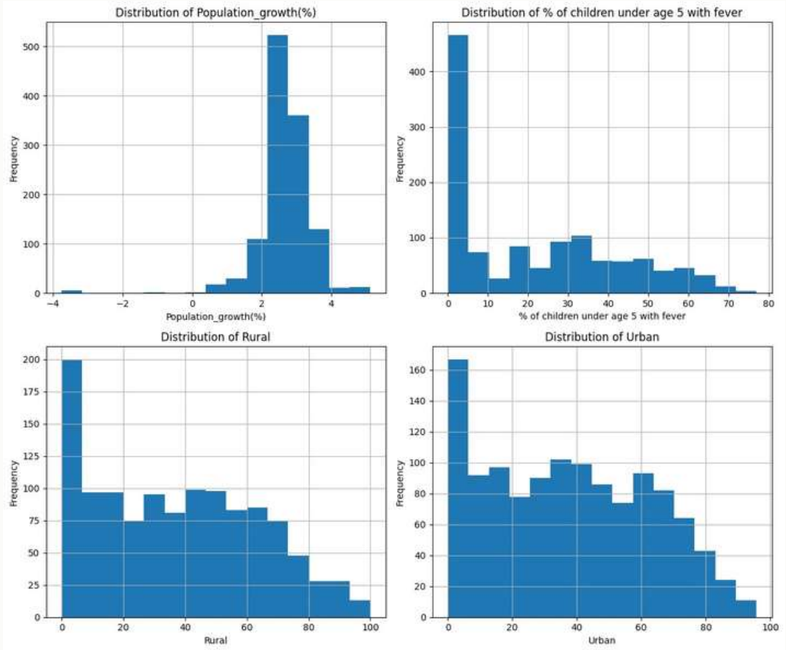
Feature Engineering:
To define our target variable, we assign weights to three critical variables that significantly influence our forecasts.

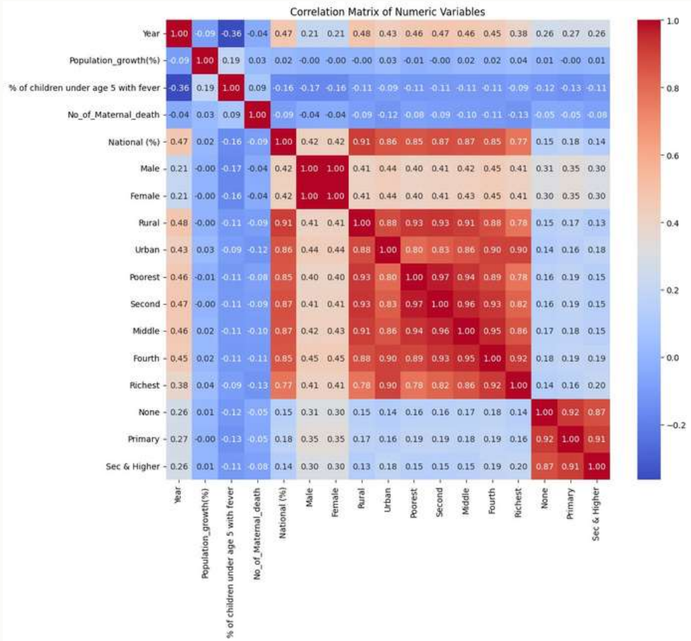
05

Model Selection, Training and Evaluation:
We apply various regression models to determine which one provides the highest accuracy for our forecasts. The models we evaluate include: Ridge Regression, Lasso Regression, ElasticNet Regression, K-Nearest Neighbors (KNN), Support Vector, Regression (SVR), Decision Tree, RandomForestRegressor

Data Model

The dataset, sourced from Hamoye and supplemented with additional relevant data, was divided into feature variables (X) and target variables (Y). To ensure the model's validity and reliability, the data was split into training and testing sets using an 80-20 ratio. This split was facilitated by the train_test_split function from the sklearn.model_selection module, with a fixed random state of 42 to guarantee reproducibility.





Model Training and Evaluation

Each model was trained on the training dataset (X_train and Y_train) and subsequently tested on the testing dataset (X_test and Y_test). The primary metric used to evaluate the models was the R-squared value.

It is evident that in both rural and urban areas of Africa, younger individuals are more prone to malaria. Consequently, the visuals show a positive skew. This indicates that special care must be taken for younger people, as they are at a significantly higher risk of contracting malaria.

This illustrates the correlation among the numerical features. Notably, it is evident that the poorest individuals in both rural and urban areas are most adversely affected by malaria, with correlations as high as 0.93 and 0.80, respectively

Data Model

The data below presents the R-squared values for various regression models applied to our dataset. This approach allows for a comprehensive comparison of different regression models, helping to identify the most accurate model for forecasting purposes based on their R-squared values. A higher R-squared value indicates a better fit for the model.

Random Forest
0.993712

Ridge Regression
0.999988

K-Nearest Neighbors
0.921915

Decision Tree
0.996905

0.863632
ElasticNet Regression

0.975381
Lasso Regression

0.844023
Support Vector Regression

Summary

Some challenges encountered in this project included sourcing additional relevant datasets beyond those provided by Hamoye. In summary. Our data model approach combined rigorous data preparation, careful model selection, and thorough evaluation to identify the most suitable methods for forecasting antimalarial drug needs.

While Ridge Regression stood out in terms of accuracy, the final choice of model must balance accuracy with practical considerations such as interpretability and computational efficiency. This holistic approach ensures that the chosen model not only performs well on historical data but is also robust and adaptable to future needs.

However, other factors such as model interpretability, computational efficiency, and potential overfitting must be considered when selecting the final model for implementation.

HAMOYE PREMIERE PROJECT

THANK YOU

Team: Jenkins