GROUP 14

CAPSTONE PROJECT

Malaria Prediction

Introduction

Project Overview

Malaria continues to be a significant public health challenge in Kenya, with substantial efforts dedicated to its prevention and control. The Kenya Demographic and Health Survey's Malaria Indicator Survey (KDHS-MIS) provides a rich source of data that, we analyzed using advanced machine learning (ML) techniques, so as to offer valuable insights and understanding of malaria prevalence, risk factors, and intervention effectiveness.

Problem Statement

The problem revolves around **the underutilization of machine learning techniques in analyzing Kenya's Demographic and Health Survey's Malaria Indicator Survey (KDHS-MIS) data,** hindering comprehensive understanding and predictive modeling of malaria prevalence, risk factors, and control efficacy.
Hence, Integrating machine learning into KDHS-MIS analysis aims to bridge this gap, providing actionable insights for targeted interventions and enhancing malaria control strategies in Kenya.

Objectives

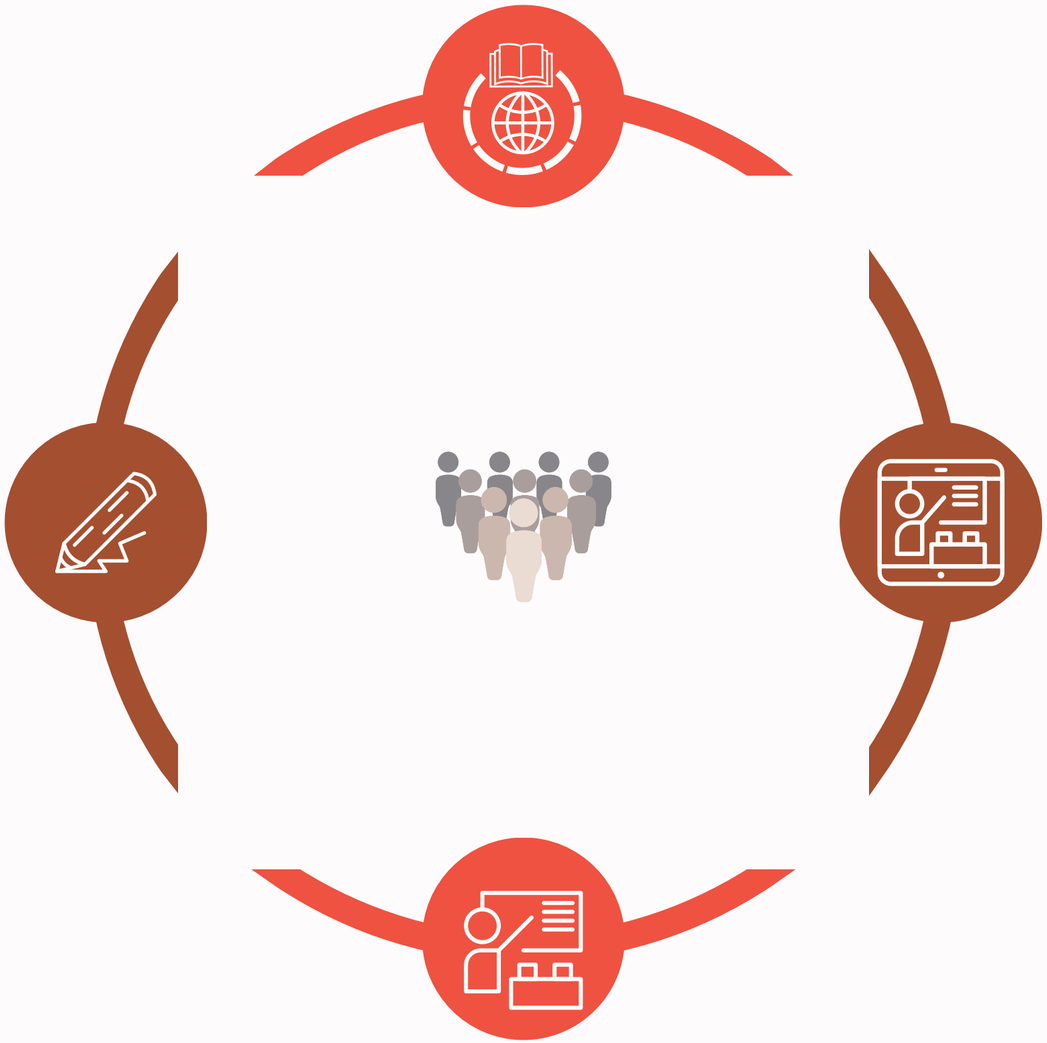
01 Apply machine learning algorithms to analyze KDHS-MIS data and identify patterns, trends, and potential predictive factors associated with malaria prevalence.

02 Develop predictive models for malaria occurrence, severity, and spatial distribution to aid in targeted intervention strategies.

03 Evaluate the effectiveness of existing malaria control measures and recommend data-driven improvements

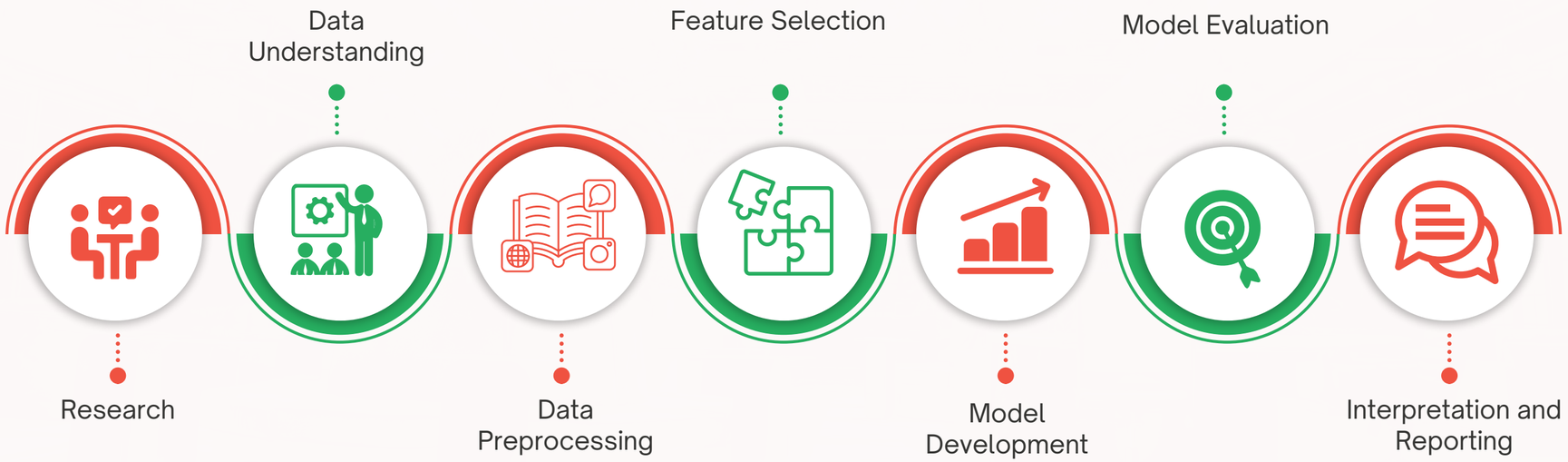
04 To Conduct comprehensive feature engineering to extract relevant features from DHS data, considering demographic, socio-economic, and health-related variables.

05 To identify regional variations in malarial - health indicators



Methodology

In this section, we use a methodical approach employed to analyze the Kenya Demographic and Health Survey-Malaria Indicator Survey (KDHS-MIS) dataset, that includes



Our Approach

Data Understanding

- Here the understanding of the data was to comprehend the nature, structure, and characteristics of the dataset.
- The data for this project was obtained from the **[DHS Program website]** (https://dhsprogram.com/data/dataset/Kenya_Standard-DHS_2022.cfm?flag=0).
- The data available is relevant for the intended analysis and predictions.

Data Preprocessing

- The KDHS-MIS dataset is **cleaned and preprocessed** to handle missing values, outliers, to guarantee data quality.
- This involved:

1. Collapsing repeated columns
2. Handling missing values by: *Filling in missing values, Dropping empty columns*
3. Handling unique values
4. Identifying columns to keep

- Techniques: (simpleImputing, threshholds, fit_transform)

Data Visualization

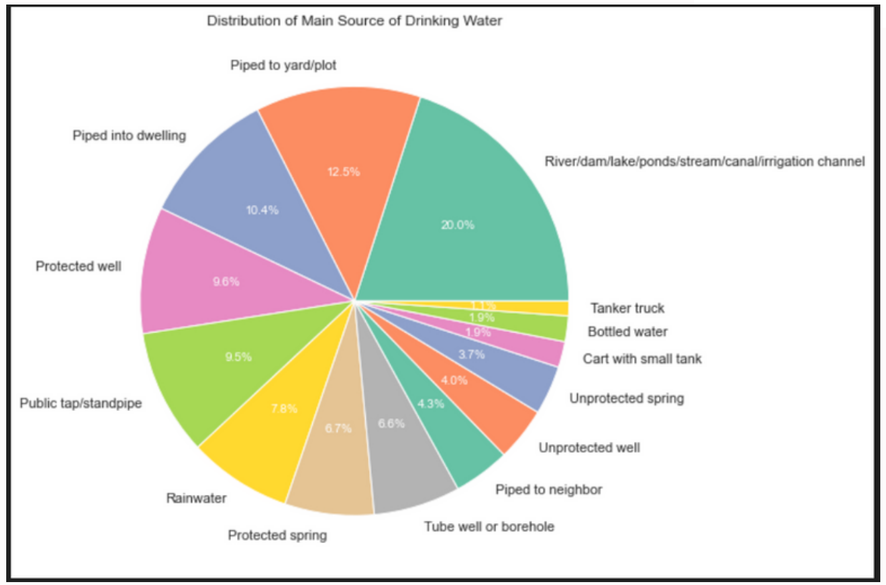
- The Exploratory Data Analysis process involved:
  1. **Univariate EDA**

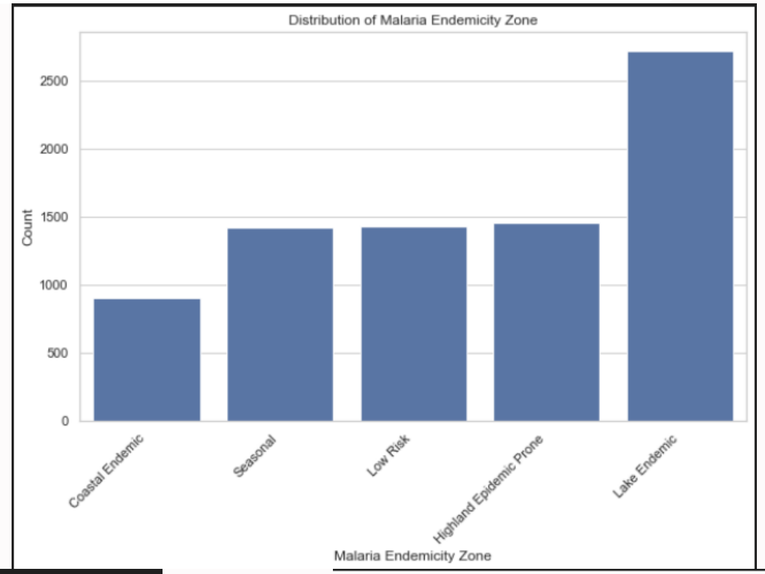
Distribution of the various aspects that affect households *(type_of_place, region_of_residence, water_sources, Number of Children, Malaria Endemicity Zone)*

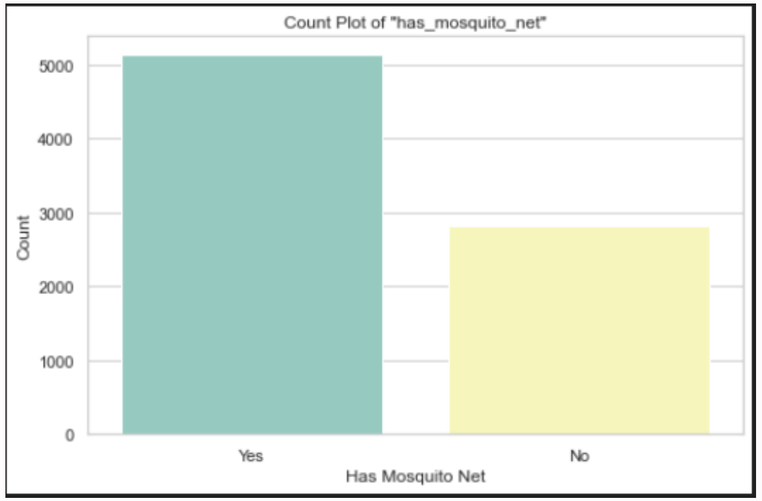
2. Multivariate EDA

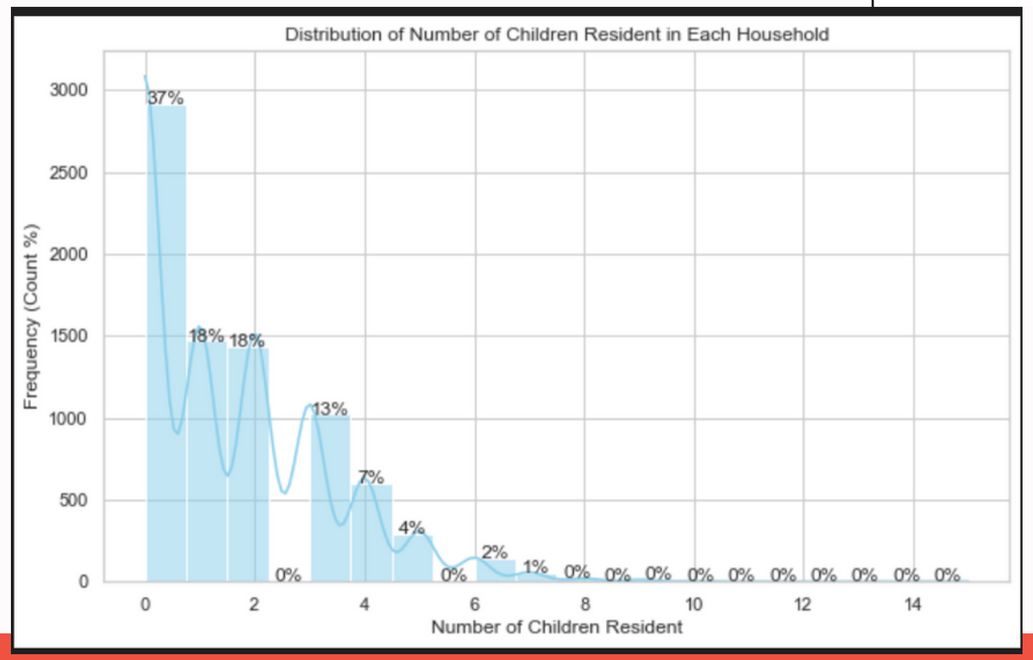
Relationship of aspects that affect households *(correlation matrix, type_of_place Vs Availability of Mosquito nets, malaria cases by region, malaria cases by age group)*

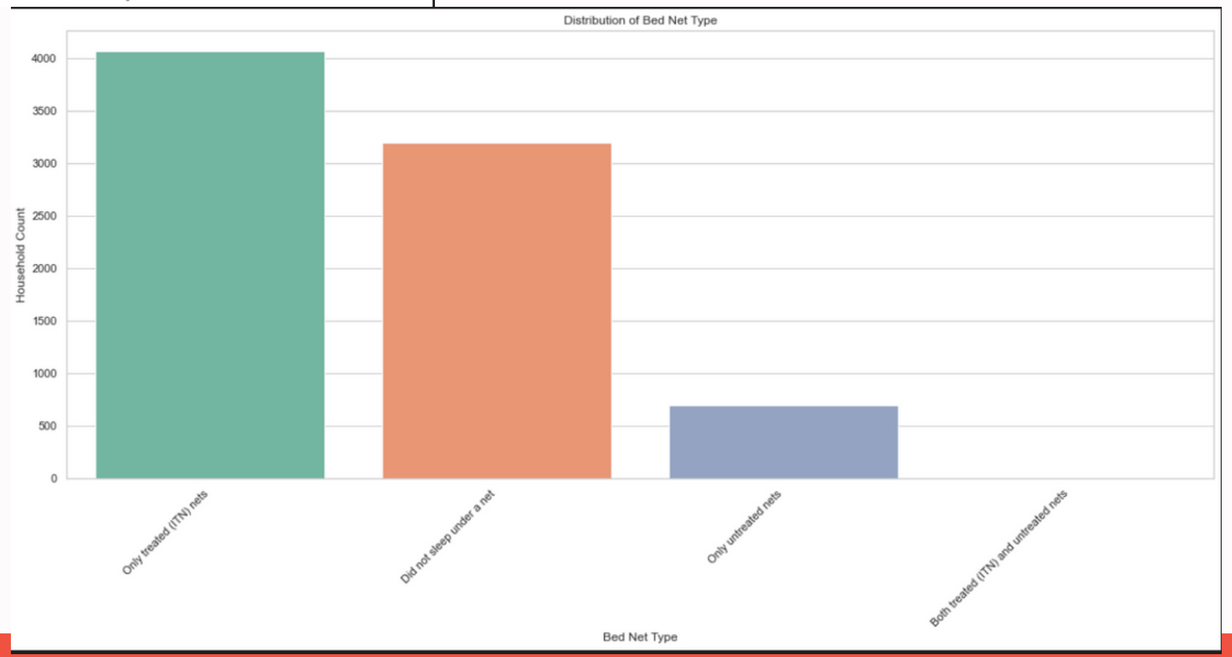
Univariate EDA



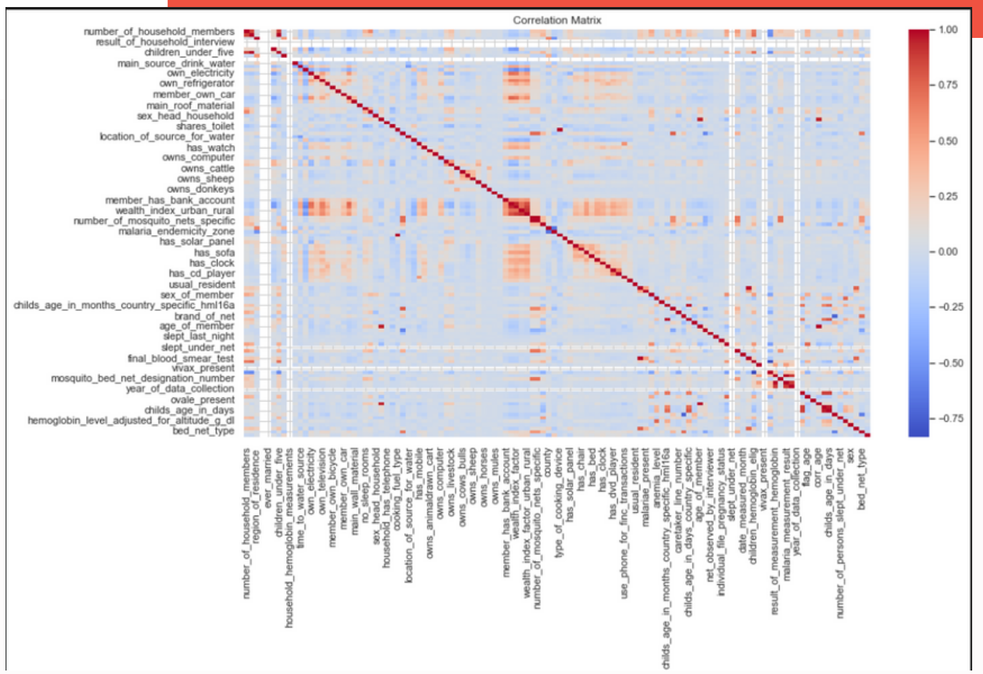


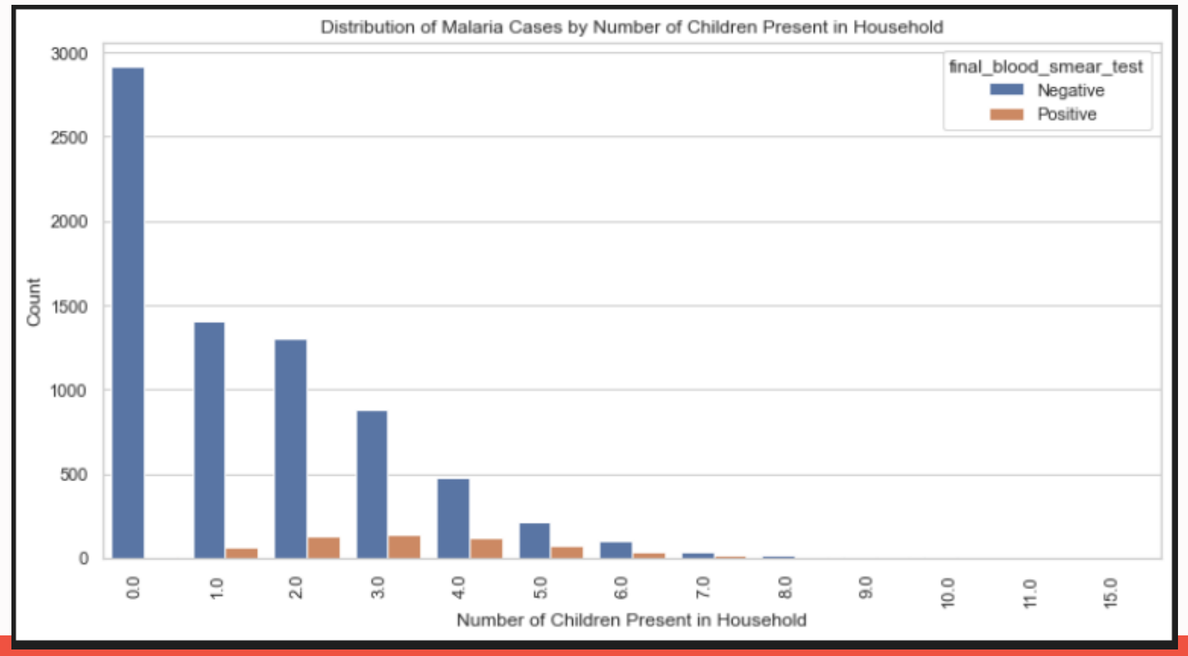


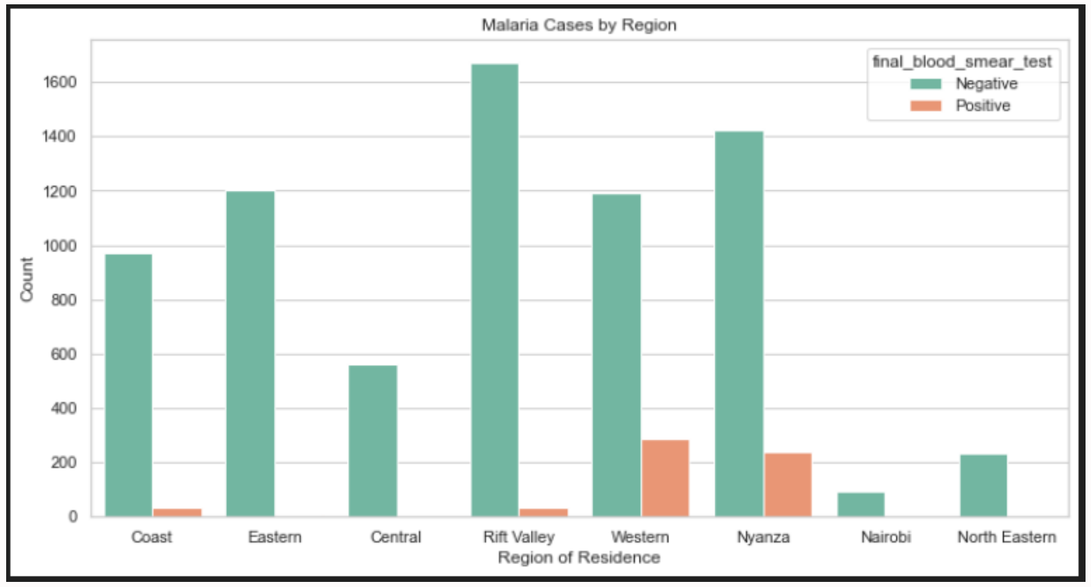




Multivariate EDA







Modelling

For our case project, We used Classification Models to predict the overall ocurrence of Malaria given the variable aspects provided

The models chosen for this dataset take into account the binary classification nature of the data, efficieny in prediction in large datasets, and the assumption of a linear relationship between the logic of the outcome variable and the predictor variables.

The models used include:

- Logistic Regression Model
- Lasso Model
- Random Forest Model

Model Evaluation

An interpretation and explanation of the key metrics:

Accuracy

Accuracy measures the overall correctness of the model predictions.

In this case, the model has an accuracy of approximately 96.67%, indicating that about 96.67% of the predictions are correct.

Precision

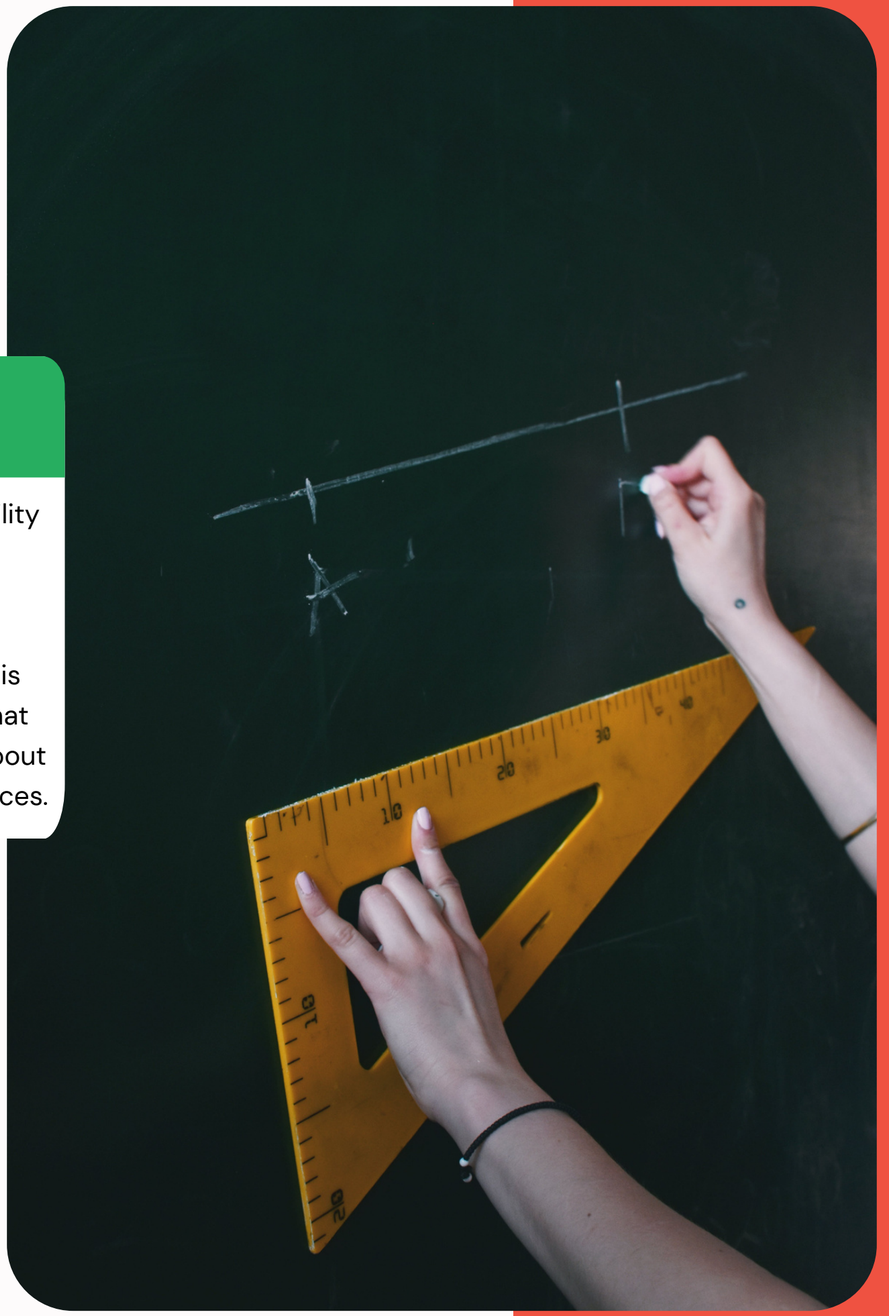
Precision measures the accuracy of the positive predictions made by the model.

Precision for positive results (1.0) is approximately 90%. It means that when the model predicts a positive result, it is correct about 90% of the time.

Recall (sensitivity)

Recall measures the model's ability to correctly identify positive instances.

Recall for positive results (1.0) is approximately 72%. It means that the model correctly identifies about 72% of the actual positive instances.



1. MCC value of 0.7794 indicates strong overall model performance, taking into account both false positives and false negatives. This metric is particularly useful in situations where there is class imbalance or where the consequences of different types of errors vary.

2. An AUC-PR of 81.86% is relatively high and suggests that the model is able to balance precision and recall well. It signifies that the model has a good ability to correctly classify positive instances while minimizing false positives and false negatives.

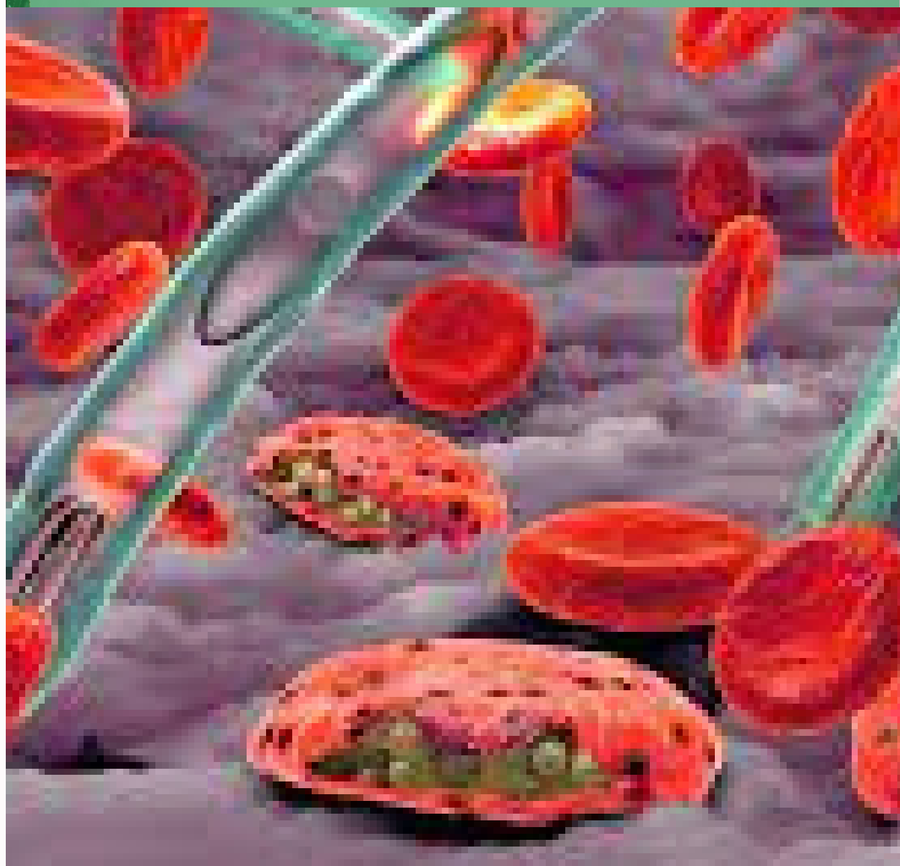
REPORTING

Recommendation

- The model can be better improved by use of hyperparameter tunning on large dataset to predict the risk of malaria prevalence and inturn inform where malaria control strategies need to be further enhanced to reduce the burden of malaria disease.

- Other than that, it employs good predictability to areas that:

Need better health care support
More distribution of Treated Mosquito Nets
Availability of better housing facilities (toilets, food preparations areas, source of drinking water)



Conclusion

The model l has a good ability to correctly classify positive instances while minimizing false positives and false negatives. Also, provides the user with high accuracy in prediction. Hence it can be used in Prediction of Malaria risk areas/indications, for strategic planning and decision making to the caise of action to be taken

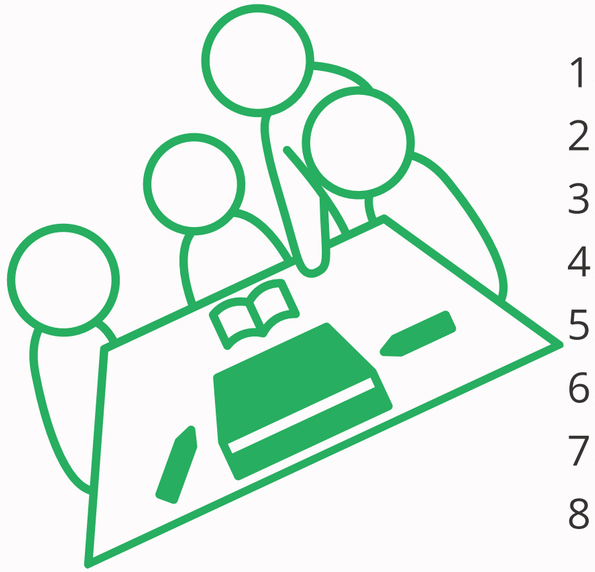
DEPLOYMENT

- We were able to user Streamlit to depoy our model

Future Work

- Visualization ofregional variations in malaria risk indicators using; heatmaps, gmap
- Further analysis to find the relationship between the malaria indicators and most common diseases in the different demographics.
- Analyze trends in malaria prevalence over time.
- Better improve the model deployed features for better prediction results (incorporating categorical variables in the user inputs)

OUR TEAM



1. **Alpha Guya**
2. **Ben Ochoro**
3. **Caleb Ochieng**
4. **Christine Mukiri**
5. **Dominic Muli**
6. **Frank Mandele**
7. **Jacquline Tulinye**
8. **Lesley Wanjiku**

THANK YOU