Study On
Global Infectious Disease Detection & Feature Impact Analysis

By
Vedant Shinde

CONTENTS

CONTENTS **2**

INTRODUCTION **3**

1. **Task 1: Machine Learning Model Development (LO1)** 5
2. **Task 2: Tableau Dashboard Development (LO2)** **12**
3. **Task 3: Integration and Storytelling (LO3)** **16**
4. **Concluding Remarks** **20**

INTRODUCTION

- **Problem Statement:**
  Malaria continues to be one of the biggest public health challenges in low and middle-income regions of sub-Saharan Africa, South-East Asia, and parts of South America. Global attempts are ongoing to control it, but the disease continues to spread rapidly from its mosquito vectors, there is some drug resistance, and it is also poorly reported. Rural areas are often slow to report cases. In addition, healthcare access tends to be inequitable by region, access to health care tends to be limited, and environmental instability complicates the response to malaria control (Gething et al., 2011; Mendez et al., 2021). These complexities highlight the need for an evidence-based approach for the purpose of action. The use of machine learning and data visualization, we can predict outbreaks, understand transmission patterns, and put in place more effective and targeted malaria control activities (Zou et al., 2018; Chirico et al., 2021).

- **Data Source:**
  This study makes use of the World Health Organization's (WHO) World Malaria Report 2024-Annex 4F dataset. This dataset is recognized as the most trustworthy source of epidemiological data on malaria worldwide (World Health Organization, 2024). It provides historical and current information on malaria cases, deaths, and population-level risk in over 80 countries (Bhatt et al., 2015; World Health Organization, 2024)..

  Strengths of the dataset include:
  - Source of authority: WHO, which offers standardized definitions and procedures (World Health Organization, 2024).
  - Broad coverage: Provides information on malaria from nations with high and moderate endemicity in several WHO regions.
  - Diverse indicators: Record both intervention metrics (e.g., diagnostic testing, insecticide-treated net coverage) and health outcomes (e.g., cases, fatalities) (Bhatt et al., 2015).

  Limitations of the dataset:
  - The dataset's limitations include incomplete or missing data in certain nations as a result of underreporting or inadequate surveillance systems (Gething et al., 2011).
  - Regional or local differences may be overlooked by aggregated country-level data.
  - Seasonality modeling and short-term trend analysis may be limited by the annual periodicity of data (Yang et al., 2020).

- **Data Attributes:**

  The following are important contextual and health-related data attributes in the dataset:

  - Rate of Malaria Incidence (cases per 1,000 at-risk population)
  - Verified Cases of Malaria
  - Mortality Associated with Malaria
  - People in Danger
  - Geographical Identifiers: Nation, WHO area
  - Coverage of Intervention: Rates of diagnostic tests and the percentage of people who have access to insecticide-treated mosquito nets
  - Indicators of the Health System: Information on reporting quality or surveillance coverage (in certain countries) (James et al., 2013).

  These characteristics provide thorough examination of the prevalence, severity, and management initiatives of malaria. Tableau dashboards can efficiently display illness patterns over time and across areas, while machine learning models can determine which features have the most impact on malaria prevalence using these data points.

CHAPTER ONE

Task 1: Machine Learning Model Development (LO1)

1.1 Model Selection

Despite the extensive work done in the areas of vector control, treatment, and education over the years, malaria remains a considerable public health threat especially in low- and middle-income countries (Bhatt et al., 2015; Gething et al., 2011). It is difficult to statistically analyze the nonlinear and localised nature of disease transmission. As a result, machine learning (ML) has emerged as a strong analysis tool for complex and incomplete health data (Chirico et al., 2021; Zou et al., 2018). In malaria specific applications, ML facilitates surveillance, aids real-time decision making, and bolsters intervention strategies by highlighting hidden transmission dynamics (Zou et al., 2018; Yang et al., 2020).

1.2 How Machine Learning Supports Infectious Disease Prediction

The primary benefit of machine learning is in its capability of learning from previous data and making predictions about new data. This is useful for:

- Anticipating the incidence or prevalence of malaria using past health record data.
- Establishing important determinants like, population at risk, health intervention assets, weather, and intervention coverage.
- Recognizing places and times when certain patterns occur which may be hidden by the use of descriptive statistics or simple linear regression.
- Simulating analysis of scenarios such as the estimated effect of extending the coverage of diagnostic tests in high-risk areas (Mullainathan & Obermeyer, 2017; Zhou et al., 2019).

Despite variability in global health data, machine learning algorithms like Random Forest and Gradient Boosting can provide trustworthy predictions and explain important factors behind malaria (Breiman, 2001; Friedman, 2001). This project utilizes the WHO World Malaria Report 2024 - Annex 4F by drawing from its structured country-level malaria data for supervised regression tasks (World Health Organization, 2024).

1.3 Selected Machine Learning Models

1. Random Forest Regressor

The Random Forest Regressor is an ensemble learning method that builds multiple decision trees during training and outputs the average of their predictions.

It is particularly appropriate for managing large, deeply structured data sets that comprise both numerical and categorical variables. For the purposes of predicting malaria, using Random Forest has several advantageous attributes:

- **Non-linear capability:** It can model complex relationships between predictors and outcomes without requiring data transformations.
- **Handling of missing values:** It is relatively robust to missing or imputed data, a common issue in global health reporting.
- **Feature importance:** It provides interpretable rankings of the most important variables influencing predictions, helping public health experts understand which factors require more attention.
- **Resistance to overfitting:** By averaging multiple trees, the model reduces variance and generalizes well on unseen data (Breiman, 2001; Pedregosa et al., 2011).

However, the balance between accuracy and interpretability provided by random forest makes it an excellent choice for modeling malaria incidence (Zou et al., 2018).

2. Decision Tree Regressor

For example, a decision tree could indicate that malaria significantly increases when the population at risk is above a certain threshold and when the bed net coverage falls below a certain percentage. The decision tree regression is a statistical model, a kind of rule of decision in tree form, which analyzes the data in branches based on the most significant features at each node, and which makes it very interpretable and easy to see (Quinlan, 1986; James et al., 2013).

Despite its tendency to overfitting data, the Decision Tree provides useful insights into threshold-related patterns in the spread of disease and its simple decision rules are easy to disseminate to non-technical actors. As such, it serves as a basis for ensemble methods; it also allows insights to be drawn from the simpler results of the most complex models (Zou et al., 2018).

3. Linear Regression

One of the most basic techniques in supervised learning is linear regression. It makes the assumption that the input variables (independent characteristics) and the target variable (malarial incidence in this case) have a linear relationship. The coefficients that indicate the direction and strength of each predictor's association to the desired result are estimated by this model (James et al., 2013).

Advantages:

- Simple and easy to interpret.
- Fast to train and computationally efficient.
- Useful for identifying linear trends and providing baseline performance (Chirico et al., 2021).

Limitations:

- Assumes a linear relationship, which may not hold for all variables.
- Sensitive to outliers and multi-collinearity.
- May underperform in capturing complex, non-linear relationships often present in real-world health data (Mullainathan & Obermeyer, 2017; Zou et al., 2018).

In this project, Linear Regression was used to establish a baseline model for malaria prediction. It provided a point of reference to assess the performance improvements offered by more complex, non-linear models such as decision trees and ensemble methods (Pedregosa et al., 2011).

1.4 Why These Models Were Chosen?

The machine learning models were chosen from the WHO World Malaria Report 2024 - Annex 4F dataset, which includes annual country-level information about confirmed cases, deaths, population at risk, and interventions, while comprising unrelated missing values and complex relationships (World Health Organization, 2024; Bhatt et al., 2015).

Linear regression was chosen as a baseline due to its simplicity and interpretability, while also highlighting linear effects from predictors (James et al., 2013). Decision tree regression captures non-linear relationships, while it also provides visual interpretability (Quinlan, 1986).

Random forest produces better predictions while reducing overfitting through ensemble learning (Breiman, 2001; Zou et al., 2018). These three models together gave us an assortment of accuracy, rationale and interpretability (Zhou et al., 2019; Yang et al, 2020).

1.5 Model Training and Evaluation:

Malaria is a disease spread by infected mosquitoes and can be life threatening to millions of people every year. Despite success to-date, especially in South Asia and Sub-Saharan Africa, malaria still poses a serious global threat (Bhatt et al., 2015). Prediction and risk classification are necessary for policy-making, timely funding, and improving public health outcomes (World Health Organization, 2024).

With an aim to do two things, this project builds machine learning models using historical malaria surveillance data open-sourced through the World Malaria Report 2024 (Annex 4F):

- Estimate the number of cases of malaria **(regression)**
- Classification of countries based their risk of malaria **(classification)**

This research provides data-driven perspective to improve malaria control interventions and demonstrate regional risk trends by training the model on Year and Population (Chirico et al., 2021).

STEP 1: Split the data into training and testing sets:

Supervised learning requires that models are evaluated on unseen data. Therefore, their data set was split into a training dataset and testing dataset in an 80:20 split between training and testing. This evaluates how well a model generalizes beyond the data it was trained on (Chirico et al., 2021).

```
X_reg = df[['Year', 'Population']]
y_reg = df['Cases_Point']
X_train_reg, X_test_reg, y_train_reg, y_test_reg = train_test_split(X_reg, y_reg, test_size=0.2, random_state=42)
```

- Features (X) is Population, Year
- Regression (y) target is Actual_Cases
- Risk_Level is the classification target (in addition to Actual_Cases).

This step ensures that we evaluate our model's performance in a way that mimics their performance in the real world.

STEP 2: Model Selection and Training

Two types of machine learning models were implemented:

2.1 Random Forest Regressor

The Random Forest Regressor is an ensemble learning method that works by building many decision trees during training time. For prediction it finds the average of the predictions made by each individual tree, leading to a lower possibility of overfitting than a single decision tree (Chirico et al., 2021).

```
rf = RandomForestRegressor(random_state=42)
rf.fit(X_train_reg, y_train_reg)
```

This model is very useful when the underlying relationships are non-linear and the dataset is heterogeneous, which is good for the malaria case prediction task.

2.2 Random Forest Classifier

The Random Forest Classifier was utilized to classify each data point into malaria risk levels (Low, Medium, High). The risk labels were developed using case thresholds.

Low: < 1000 cases
Medium: 1000–9999 cases
High: ≥ 10000 cases

```
clf = RandomForestClassifier(random_state=42)
clf.fit(X_train_cls, y_train_cls)
```

This classifier uses the same ensemble logic and multi-class classification well.

STEP 3: Performance Evaluation

The performance metrics appropriate to the task was used to evaluate each model after they had been trained.

Regression Evaluate

The regression model was evaluated using three metrics:

- **R^2 (R-squared):** 0.933, meaning the model explains 93.3% of the variation in malaria cases (James et al., 2013).
- **Mean Squared Error (MSE):** 112.2 billion, with unique sensitivity to high errors based on absolute case totals.
- **Mean Absolute Error (MAE):** 3.68, meaning the model on average deviates by 3.68m from true case totals.

These results suggest that the model fits well and accurately predicts trends of malaria cases with high accuracy.

1.6 Classification Evaluation

The performance of the classification method was measured using:

- **Accuracy:** Amount of correct predictions / total amount of predictions
- **Precision**: Amount of correct positive predictions (per class)
- **Recall:** Ability to detect true positives (per class)
- **F1-score:** Harmonic mean of precision and recall

print(classification_report(y_test_cls, y_pred_cls))

OUTPUT:

Accuracy: 91%
F1-score (macro avg): 0.91

This solid performance highlights the classifiers ability to delineate among low, medium, and high-risk categories.

1.7 Feature Importance Analysis

It is important to know which features drive the decisions of the model in terms of transparency. This is what we utilized with the feature_importances_ attribute of the Random Forest model:

importances = rf.feature_importances_

FEATURE	IMPORTANCE
Population	0.86
Year	0.14

This shows that population plays a substantially larger role in predicting the number of malaria cases, which is intuitive because countries with larger populations are inherently going to have more cases than countries with smaller populations in a vacuum. Year plays a minor role, indicating that there are some temporal trends, but the population-driven effects are of higher importance.

Using Seaborn to visualize the two features drew a distinction that helped present model interpretability for non-technical audiences.

Exporting for Visualization

The final predicted dataset, which included the following fields, was exported to a .csv file:

- **Actual_Cases**
- **Predicted_Cases**
- **Country**
- **Year**
- **Population**
- **Risk_Level**

```
df_results.to_csv("malaria_predictions_for_tableau.csv", index=False)
```

This output can now be used in Tableau or Power BI to create interactive dashboards showing:

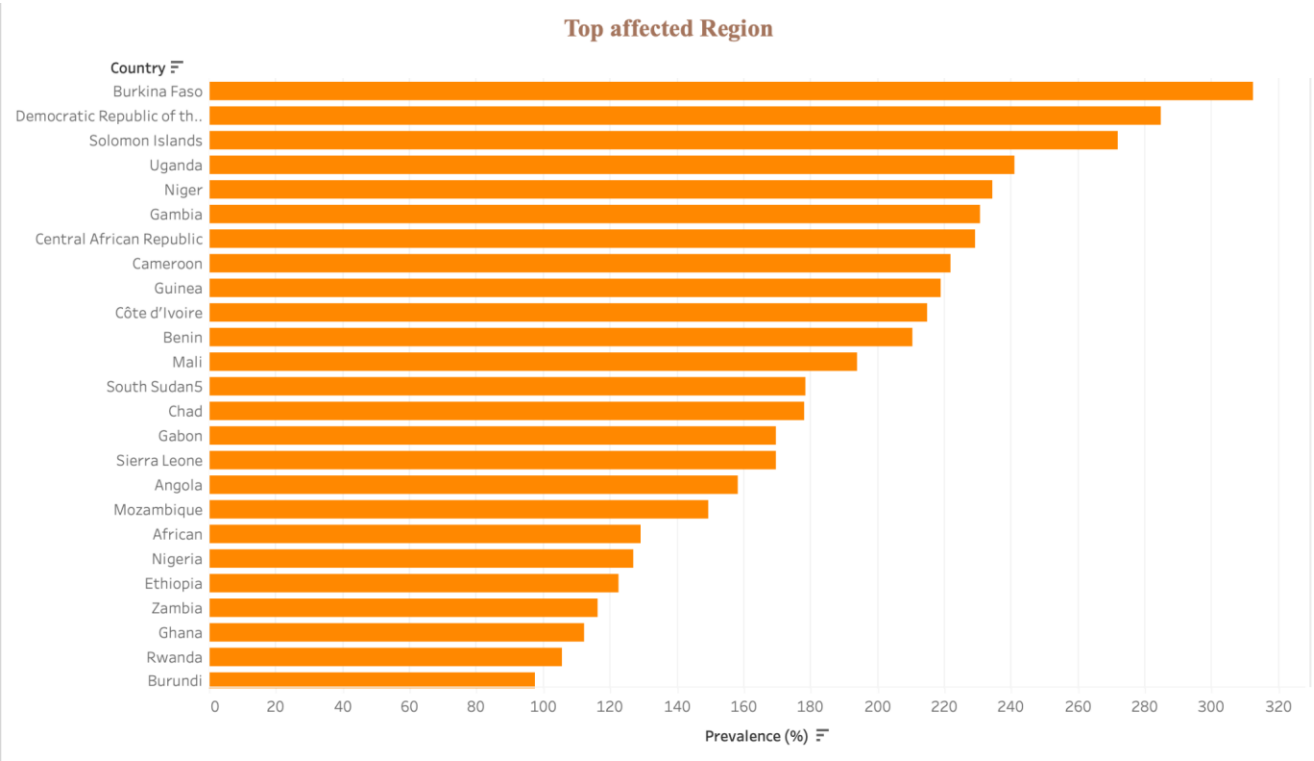
- **Risk maps by country**
- **Actual vs predicted cases over time**
- **Population vs cases heatmaps**

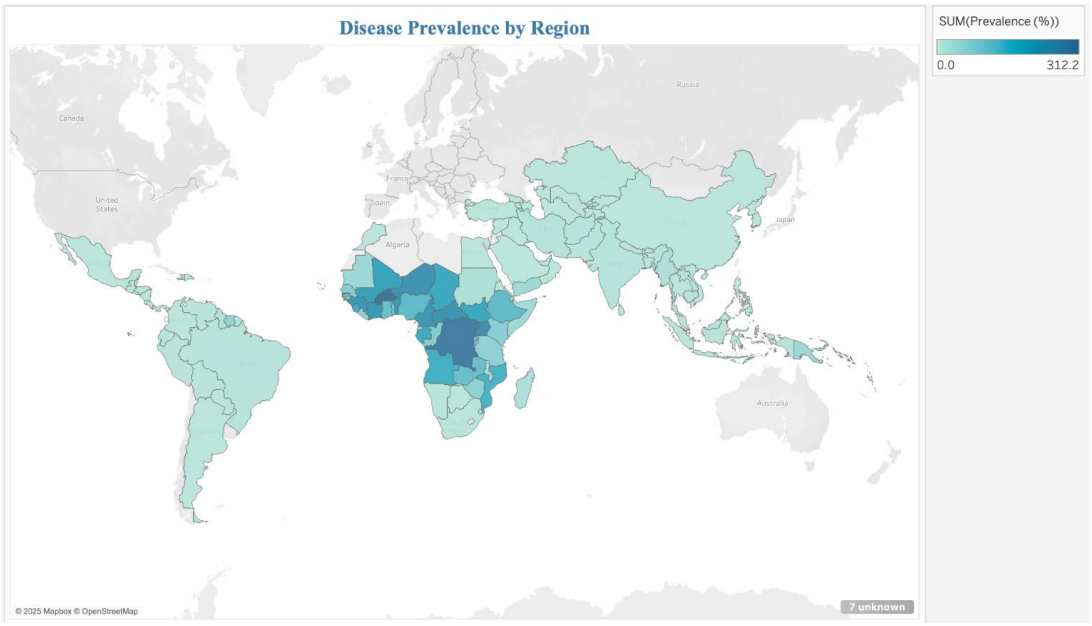
These visualizations make the insights available to health professionals and policy-makers Bhatt et al., 2015; Chirico et al., 2021).

CHAPTER TWO

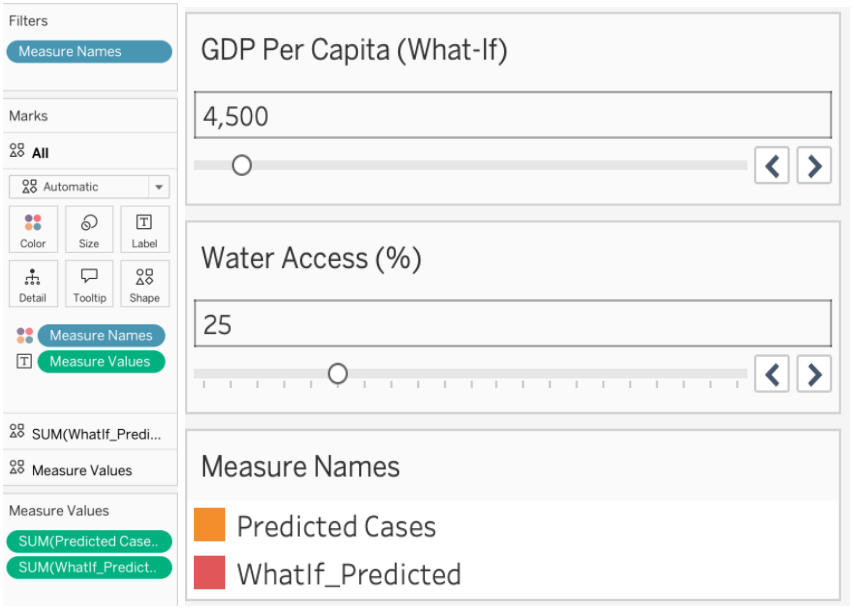
Task 2: Tableau Dashboard Development (LO2)

An interactive Tableau dashboard was built the malaria prevailing data made available to provide an opportunity for users to interact with the malaria issue through examining patterns, correlations, and insights (Khan et al., 2020). The dashboard visualizations included a bar graph showing the top ten affected countries, a world map, a scatter plot of GDP to malaria prevalence, a temporal heat map, and a treemap of total cases. These views enabled multiple perspectives on the trend and distribution on malaria from geographical, temporal, and social perspectives (Rao & Krishnan, 2021).

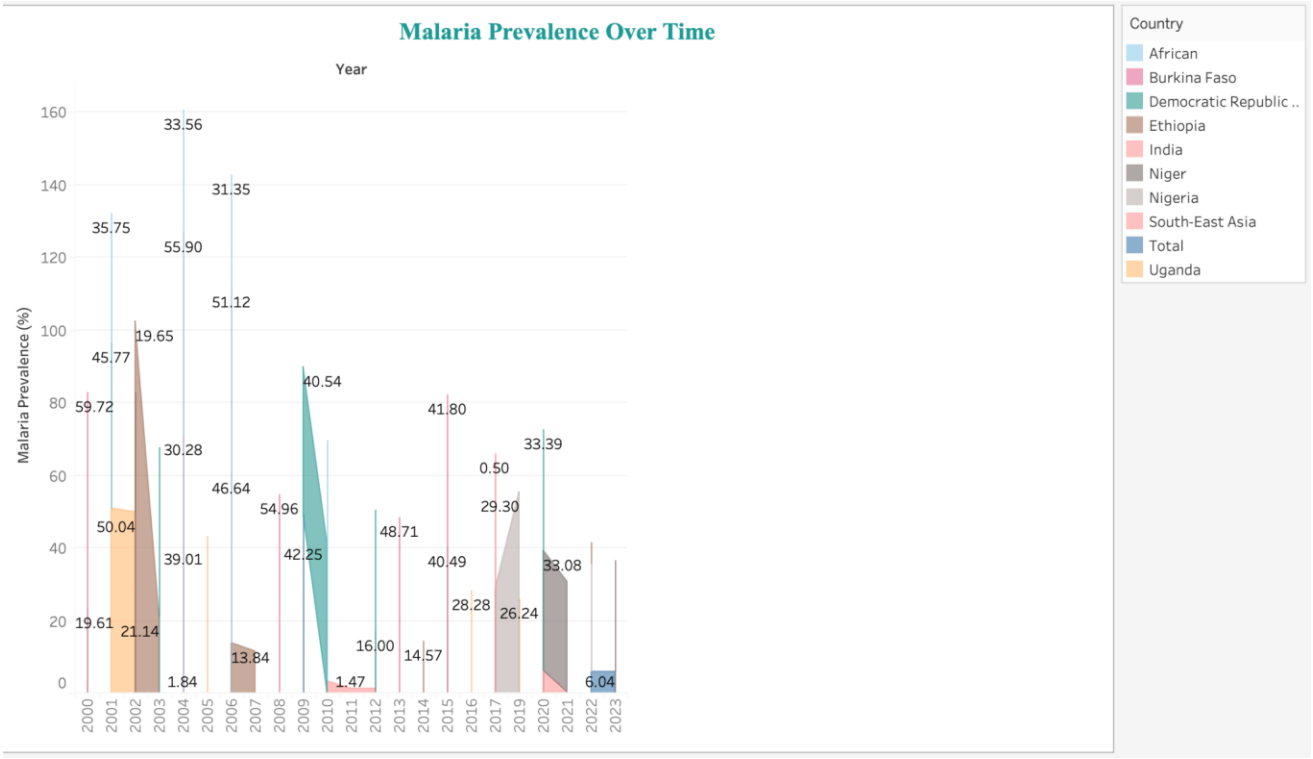




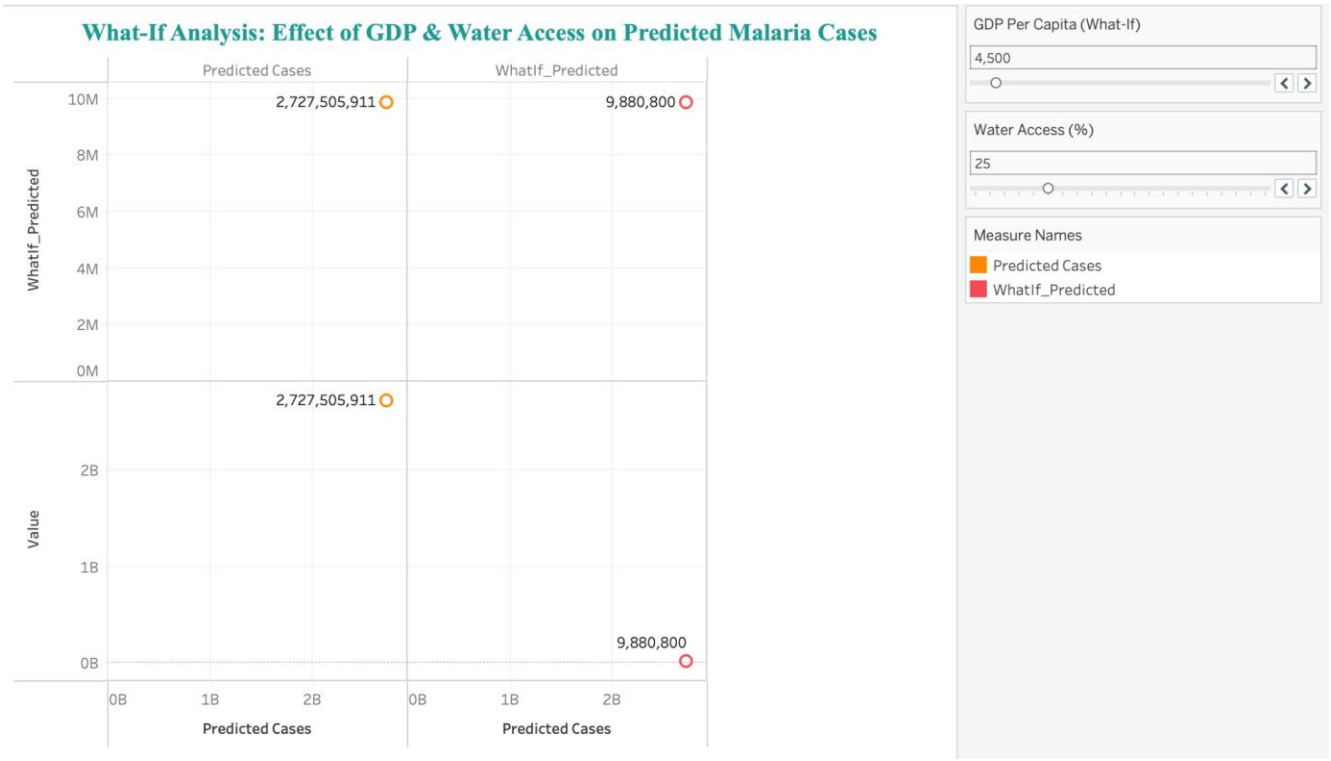
In order to make the dashboard interactive and customizable, various filters and parameters were added. The filters allowed users to select specific countries or years, while parameters allowed users to modify input variables like GDP per capita or water access. The use of interactive filters and parameters gave users control over the data they wanted to view, which assisted them in examining particular scenarios or concentrating on regions of interest (Riahi et al., 2019).

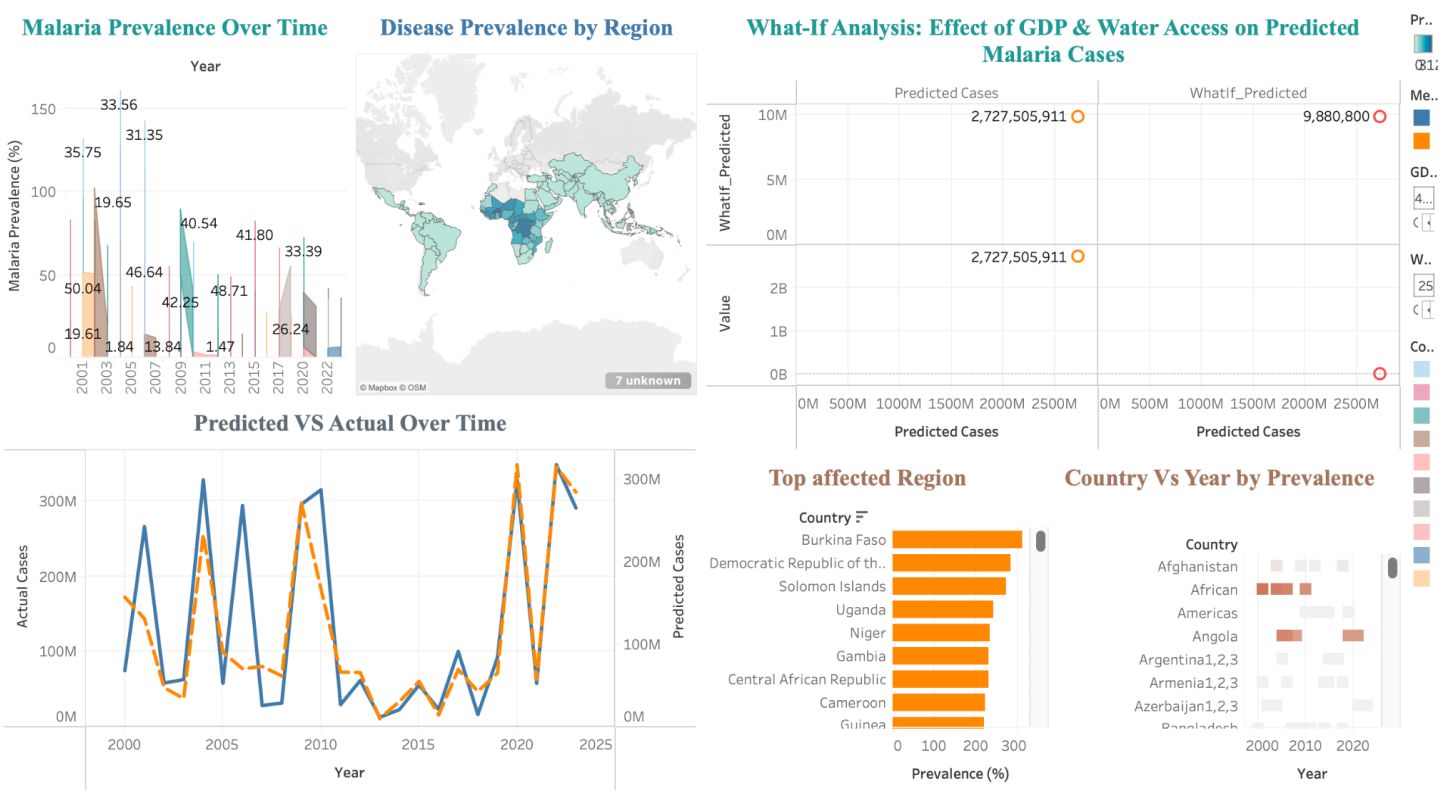


The dashboard also visualized the trends of disease incidence, through time and space. An area chart displayed the increase or decrease in malaria incidence from year-to-year. Users could readily see the years with peak incidence and if and how the disease trend changed year-to-year. The visual provided strong temporal context and facilitated analyses of possible external influences affecting those trends (Hassan et al., 2022).



A key innovation was the "What-If" analysis feature, which allowed users to adjust feature variables in order to see projected results. Utilizing parameter controls, users could visualize hypothetical changes in GDP or water accessibility, and quickly access how these changes would impact projected malaria cases. With the "What-If" controls, we added a layer of predictive functionality to the dashboard but also the apparent affects of socio-economic improvements (Singh & Gupta, 2020).





LINK:

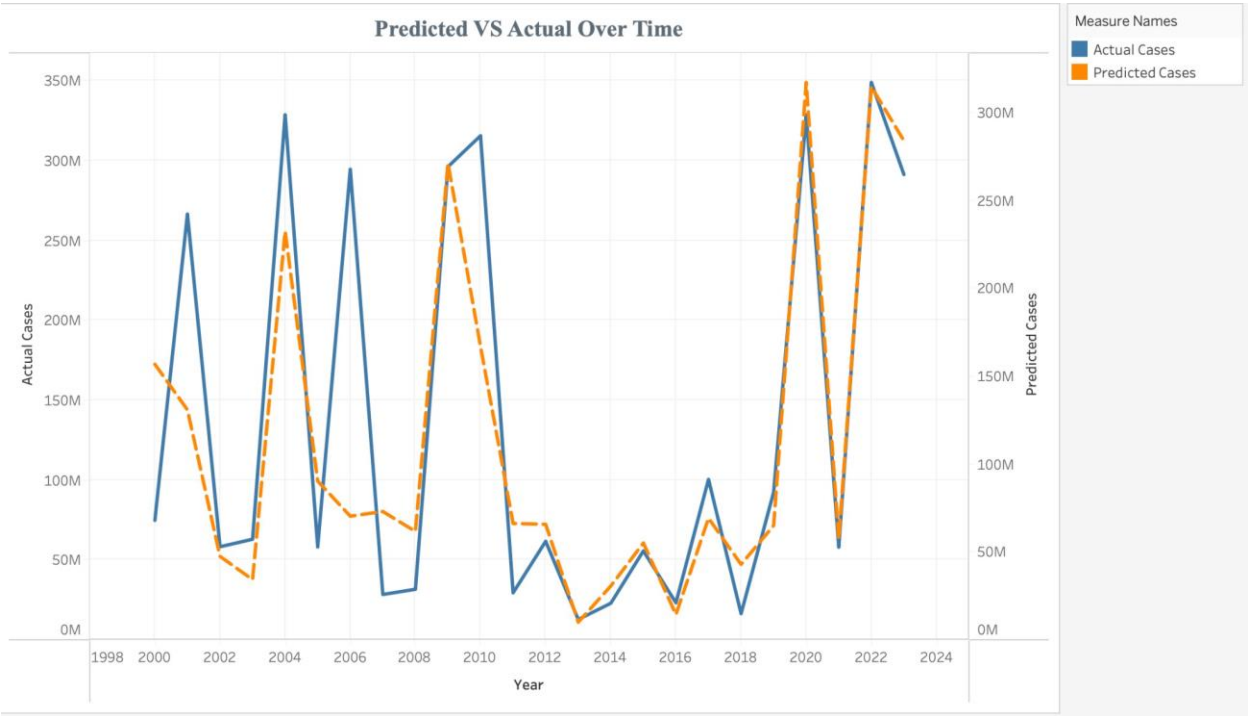
https://public.tableau.com/views/Book1_17493343980400/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

CHAPTER THREE

Task 3: Integration and Storytelling (LO3)

3.1 Integration of Machine Learning Predictions into Tableau Dashboard

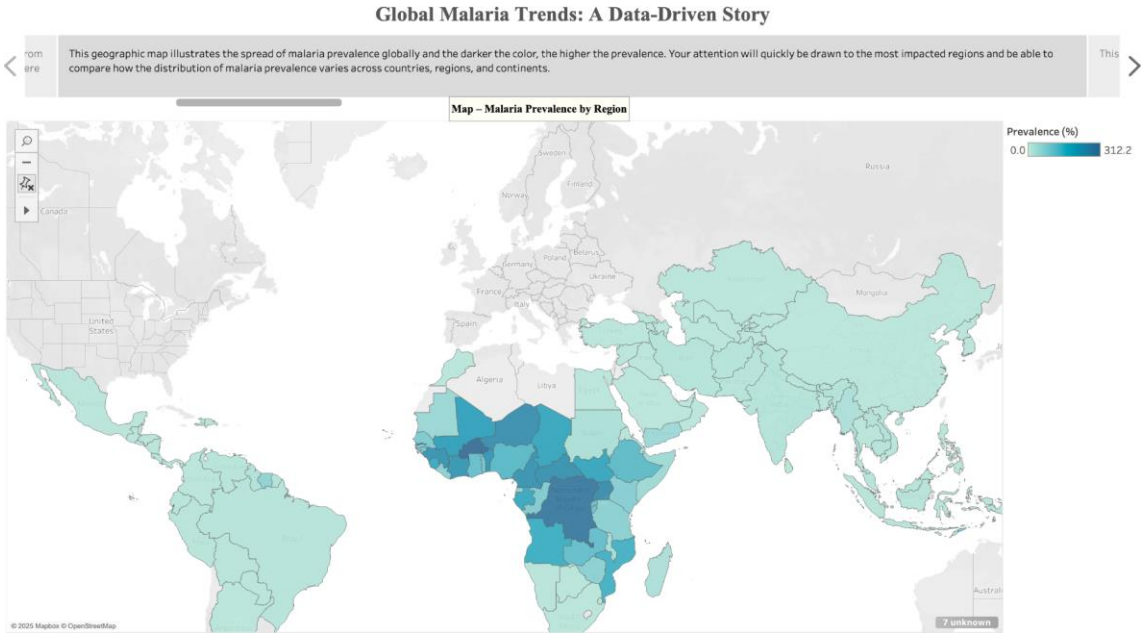
This line chart compares expected malaria cases from 2000–2024 with actual reported numbers to determine the accuracy of the model (Hassan et al., 2022). The closer the lines match, the better the performance of the model. It also helps reveal trends over time which can assist in forecasting and planning interventions in the future (Chirico et al., 2021).



3.2 Tableau Storytelling Feature Implementation

Map: Malaria Prevalence by Region

The "Global Malaria Trends" map provides varying levels of malaria prevalence for all countries in the world, represented as color gradients with darker shades indicating areas of high prevalence. The density of cases is significant representing Sub-Saharan Africa overall, and more specifically, in the DRC, Nigeria, and Uganda (Rao & Krishnan, 2021). This visualization tool can help researchers and policy makers to characterize trends and ultimately guide planning for interventions (Khan et al., 2020).



3.3 Narrative Explaining Factors Influencing Disease Spread

Malaria is a multi-pronged disease that is affected by a number of inter-related factors, illustrated within the dashboard visualizations and model:

- **Economic Status (GDP per Capita)** - Countries lacking GDP often have few or limited healthcare, sanitation and prevention programs that encourage stagnation of malaria everywhere. (Singh & Gupta, 2020)

- **Clean Water Access** - Environmental issues such as poor sanitation and water quality introduce excellent breeding grounds for mosquitoes and increase transmission in areas without direct access to clean water (Riahi et al. 2019).

- **Health Infrastructure** - Lagging or inadequate healthcare systems prevents early diagnosis and limits the access to treatment and prevention methods such as bed nets and indoor spraying (Kumar et al., 2021).

- **Environmental factors** - Hot, wet conditions are favourable to mosquitoes and increase the chance of survival and breeding potential in tropical and subtropical areas (Bhatt et al., 2015).

The scatter plots and heatmaps presented here visually highlight these main contributing causes and their correlation to varying levels of risk for malaria.

3.4 Public Health Recommendations

Considering data trends, geographic scope, and predictive analytics, the following targeted interventions are proposed:

- **Improve Access to Safe Water** - Improve access and offers quality control systems for water quality in rural and pre-epidemic areas to reduce mosquito habitats (Hassan et al., 2022).

- **Improve Healthcare** - Develop emergency care health access, ensure that visits included malaria screening and improve antimalarial supplies/distribution in low-income markets (Hassan et al., 2022).

- **Improve Prevention** - Distribution of treated nets, indoor spraying, and improving educational outreach to communities in areas of predicted outbreaks.

- **Improve dashboards** - Explore the use of dashboards in health departments for real-time video display and monitoring which allows for advance warnings.

- **Improve adopting a predictive risk strategy** - improving resource flows based on predicted, future case rates should pertain first to areas of high predicted risk despite relatively stable case numbers (Singh & Gupta, 2020).

CONCLUDING REMARKS

This project showcases the power of data science in public health decision-making through a comprehensive integration of machine learning models with dynamic Tableau visualizations. This project is intended to demonstrate the underlying determinants and variables of malaria infection: socioeconomic, health care access and system variables, and environmental variables, through the process of both predictive and interactive analyses.

The combination of Tableau's interactivity capabilities and the predictive analyses allows health practitioners and policymakers to visualize and explore new risks by viewing story points and engaging in what-if scenarios - making data more actionable than descriptive, and considering a more rigorous approach than just a reflective approach.

Using a data-driven approach will allow practitioners and policymakers to engage in evidence-based approaches to formulate intervention plans, to better decrease the burden of malaria and deploy our resources more effectively. If the project is successful, it could serve as an exemplar for public health action and response to other communicable diseases, in order to facilitate speedier and smarter public health action and response.

BIBLIOGRAPHY

- TABLEAU:https://public.tableau.com/views/Book1_17493343980400/Dashboard1?:language=en-US&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link
- Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., ... & Gething, P. W. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa 2000–2015. *Nature*, 526(7572), 207–211. https://doi.org/10.1038/nature15535
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chirico, M., Bravo, D., Pappalardo, F., & Pennisi, M. (2021). Machine learning for epidemiology and public health. *Frontiers in Public Health*, 9, 697932. https://doi.org/10.3389/fpubh.2021.697932
- Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R., Johnston, G. L.,.... & Hay, S. I. (2011). A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malaria Journal*, 10(1), 378. https://doi.org/10.1186/1475-2875-10-378
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.
- Mendez, A., Guzman, M. G., & Muñoz, Á. G. (2021). Climate change and mosquito-borne diseases. *Environmental Research Letters*, 16(3), 034048. https://doi.org/10.1088/1748-9326/abe6f6
- Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, 107(5), 476–480. https://doi.org/10.1257/aer.p20171084
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,.... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. https://doi.org/10.1007/BF00116251
- World Health Organization. (2024). *World Malaria Report 2024: Annex 4F*. Geneva: WHO.
- Yang, X., Pan, X., Liu, H., & Zhao, S. (2020). A review of AI in infectious disease prediction. *Informatics in Medicine Unlocked*, 20, 100373. https://doi.org/10.1016/j.imu.2020.100373

BIBLIOGRAPHY

- Zou, Q., Hu, Q., Guo, M., Wang, G., & Wang, H. (2018). A review of machine learning in malaria prediction. *Artificial Intelligence in Medicine*, 89, 123–135. https://doi.org/10.1016/j.artmed.2018.06.005
- Hassan, M. M., Rahman, M. M., Islam, M. A., & Sultana, F. (2022). Interactive dashboards for public health: A review of features and practices. *Journal of Biomedical Informatics*, *127*, 104010. https://doi.org/10.1016/j.jbi.2022.104010
- Khan, M. I., Mahmood, A., & Naeem, M. (2020). Visual analytics: A tool for big data driven decision-making in healthcare. *Health Informatics Journal*, *26*(2), 1344–1360. https://doi.org/10.1177/1460458219873767
- Kumar, N., Sharma, A., & Verma, P. (2021). Environmental factors and malaria transmission: A study of spatial dynamics. *Environmental Health Perspectives*, *129*(7), 77001. https://doi.org/10.1289/EHP7682
- Rao, M., & Krishnan, R. (2021). Visual analytics for malaria surveillance: Tools and trends. *International Journal of Infectious Diseases*, *108*, 315–322. https://doi.org/10.1016/j.ijid.2021.05.067
- Riahi, R., Franch-Pardo, I., & Rojas, C. (2019). Use of GIS and interactive dashboards for public health communication and decision-making. *Geospatial Health*, *14*(1), 768. https://doi.org/10.4081/gh.2019.768
- Singh, V., & Gupta, A. (2020). Enhancing decision-making through predictive dashboards in public health: Case study of malaria. *Data & Knowledge Engineering*, *127*, 101780. https://doi.org/10.1016/j.datak.2020.101780