PREDICTING MALARIA OUTBREAKS IN RURAL LIBERIA USING MACHINE LEARNING

FTL Liberia Group Six (6)

ABSTRACT

This project proposes to build a machine learning model to predict malaria outbreaks in rural Liberia using environmental, demographic, and health data. The goal is to create an early warning system that helps manage resources and save lives, supporting Sustainable Development Goals (SDGs) 3 and 1.

Group Members:

1. Nehemiah Kemayah
2. Thomas Adonis Marwolo
3. Robert Bright Yekeh
4. Joel J. Barclay

Project Overview

This project aims to develop a predictive analytics system for **malaria outbreaks** in rural regions of Liberia, aligning with the Sustainable Development Goal (SDG) 3: **Good Health and Well-being**. The goal is to leverage data-driven insights to improve early warning capabilities and optimize healthcare resource allocation, ultimately reducing disease burden and supporting local malaria control efforts. The problem is the persistent, seasonal nature of malaria morbidity and deaths, especially in underserved areas, with significant social and economic costs.

Objectives

- Build and validate machine learning models to predict malaria outbreak likelihood at the district level.
- Integrate climatic and non-climatic variables for higher accuracy.
- Provide actionable insights for targeted malaria interventions.
- Improve responsiveness and impact of health services by enabling timely resource planning.

Background

Malaria remains a major health challenge in many African regions, driven by complex interactions between climate, environment, and population. While previous approaches have relied on statistical forecasting and resource-intensive traditional methods, machine learning enables scalable, context-aware predictions from heterogeneous datasets. Studies in The Gambia and elsewhere show that models like Decision Trees, Random Forests, and XGBoost effectively combine climate and clinical data to enhance outbreak forecasting, which is critical for rural resource-constrained settings.

Methodology

Key machine learning techniques to be used include:

- Supervised classification methods: **Random Forest, Decision Tree (C5.0), XGBoost, Logistic Regression, KNN, SVM, and ANN**.
- Feature engineering for climate (rainfall, temperature, humidity), population, and seasonal patterns.
- Data preprocessing: normalization, imputation for missing values, and up sampling for imbalanced classes.
- Hyperparameter tuning and cross-validation (repeated 10-fold) for robust performance assessment.

Architecture Design Diagram

Overview

- **Data Ingestion**: Clinical records, meteorological data, population statistics.
- **Preprocessing**: Cleaning, normalization, imputation, feature selection.
- **Model Training**: Pipeline for multiple algorithms, cross-validation, assembling.

- **Prediction Service**: API exposing outbreak predictions.
- **Visualization Dashboard**: Web-based summary for stakeholders and decision-makers.

Components

- Data Sources API: Integrates various datasets.
- Preprocessing Module: Ensures data quality.
- ML Models: Multiple classifiers with ensemble option.
- Evaluation Metrics: Accuracy, Sensitivity, Specificity, ROC/AUC.
- Deployment: Cloud/server-hosted, dashboard access

Data Source

The project will use historic meteorological (rainfall, temperature, humidity) and clinical malaria datasets collected monthly from local health records, supplemented by census population data. Data will be normalized to mitigate variable influence, missing values imputed by month averages, and outliers managed. This multimodal data approach is essential for accurate, localized predictions.

Literature Review

Recent literature confirms that ensemble machine learning approaches (XGBoost, RF, Decision Trees) reliably predict malaria outbreaks when integrating climate and non-climate variables at the district level. Past research also emphasizes the value of hyperparameter tuning, feature selection, and robust validation for high-impact early warning systems in under-resourced contexts.

Implementation Plan

Technology Stack

- **Programming Languages**: Python (scikit-learn, XGBoost, pandas, NumPy), R
- **Frameworks/Libraries**: scikit-learn, TensorFlow/Keras, caret (R), matplotlib, seaborn
- **Visualization**: Plotly, Dash, PowerBI for dashboarding
- **Software/Hardware**: Cloud VM/server for deployment, API server (Flask/FastAPI), basic web UI

Timeline

Stage	Week 1-2	Week 3-4	Week 5-6	Week 7-8	Week 9-10
Data Collection & Cleaning	*				
Data Preprocessing & Exploration	*	*			

Model Development & Feature Eng.		*	*		
Training & Validation			*	*	
Deployment & Visualization				*	*
Reporting & Presentation					*

Task Distribution Matrix (for two members)

Task	Member 1	Member 2
Data Gathering	Lead	Support
Preprocessing	Support	Lead
Model Building	Lead	Support
Validation & Tuning	Support	Lead
Deployment	Lead	Support
Dashboard/Reporting	Support	Lead

Milestones

- **Data readiness** (cleaned and imputed datasets)
- **Model selection and training** (algorithms benchmarked, best model chosen)
- **Validation complete** (cross-validation, performance metrics reported)
- **Deployment ready** (API and dashboard functional)
- **Final report/presentation** (documentation, stakeholder-facing deliverables)

Challenges and Mitigations

- **Data quality:** Address missing and noisy data with robust preprocessing and imputation.
- **Class imbalance:** Use up sampling techniques during training to improve model generalization.
- **Model performance:** Employ multiple models and ensemble approaches; hyperparameter tuning to optimize results.
- **Technical constraints:** Use cloud resources if local capacity is limited; modular coding for flexibility.

Ethical Considerations

- Ensure **data privacy** for health records; anonymize sensitive data sets.
- Address potential **algorithmic bias** by auditing models for fairness.
- Strive for transparency in model outputs to support responsible public health interventions.

- ❖ Create a framework for community feedback to monitor unintended impacts.