ALARIA

LIBERIA

ARNING

p 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.

Nehemiah Kemayah

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

Data Preparation/Feature Engineering

1. Overview

The data preparation and feature engineering phase forms the backbone of our malaria outbreak prediction system for rural Liberia. This phase is crucial as it transforms raw environmental, demographic, and health data into meaningful features that machine learning algorithms can effectively utilize to predict malaria outbreaks. The significance of this phase lies in its ability to capture the complex temporal and spatial relationships between climatic conditions, population characteristics, and malaria transmission patterns that are essential for accurate outbreak forecasting in resource-constrained rural settings.

2. Data Collection

Our project utilizes multiple data sources as specified in the project requirements, though we acknowledge limitations in accessing real-time Liberian health system data:

Primary Data Sources:

- **Health Data**: Originally intended from Liberia's Ministry of Health DHIS2 system (CSV format, weekly malaria case reports by region)
- **Climate Data**: World Bank Climate Change Knowledge Portal providing rainfall, temperature, and humidity data (CSV format)
- **Population Data**: Liberia Institute of Statistics demographic information (Excel/CSV format)

Data Access Reality: Due to limited access to live DHIS2 and LMIS systems mentioned in our project files, we generated realistic synthetic data that follows the documented patterns from our literature review. This synthetic data maintains the same statistical properties, seasonal patterns, and climate-health relationships identified in studies from neighboring West African countries like The Gambia and Burkina Faso.

Preprocessing Steps During Collection:

- Temporal alignment of datasets to monthly frequency
- Geographic standardization using district-level aggregation
- Data validation against known seasonal malaria patterns
- Quality checks for missing values and outliers

3. Data Cleaning

Our data cleaning process addresses the fragmented reporting and data quality issues highlighted in our project documentation:

Missing Value Handling:

- Monthly average imputation for climate variables (consistent with project methodology)
- Forward-fill and backward-fill techniques for time series continuity
- Median imputation for population density data where census information was sparse

Outlier Detection and Treatment:

- Z-score method (threshold ±3) for extreme climate readings that could indicate sensor errors
- Domain knowledge-based thresholds: rainfall >500mm/month, temperature outside 15-40°C range
- Winsorization at 5th and 95th percentiles for population density to handle urban/rural extremes
- Clinical validation for malaria case counts exceeding 200 per 1000 population

Data Quality Assurance:

- Temporal consistency checks ensuring no future-dated records
- Seasonal pattern validation against literature-documented trends
- Cross-referencing climate data with regional meteorological patterns
- Healthcare facility reporting completeness assessment

4. Exploratory Data Analysis (EDA)

Our EDA reveals critical insights that inform both feature engineering and model development:

Temporal Patterns Discovered:

- Clear seasonal malaria peaks during May-October rainy season (75% higher incidence)
- Consistent 2-3 month lag between rainfall peaks and malaria outbreak spikes
- Temperature sweet spot of 20-30°C showing highest transmission rates
- Year-over-year consistency in seasonal patterns across all districts

Spatial Distribution Insights:

- Montserrado and Nimba districts showing 40% higher baseline outbreak rates
- Strong correlation (r=0.68) between population density and outbreak severity
- Rural areas with poor drainage showing elevated risk during wet seasons
- Coastal vs inland variations in humidity-outbreak relationships

Climate-Health Correlations:

- Rainfall-malaria incidence: r=0.72 (strongest predictor)
- Temperature-outbreak probability: Inverted U-shaped relationship
- Humidity optimal range: 60-90% for sustained transmission
- Compound climate events (high rainfall + optimal temperature) increasing outbreak probability by 85%

Key Statistical Findings:

- Overall outbreak rate: 32% of district-months classified as outbreaks
- Average cases during outbreaks: 67.3 per 1000 vs 8.2 per 1000 during non-outbreak periods
- Seasonal variation: 3.2x higher outbreak probability in rainy vs dry season
- District heterogeneity: Outbreak rates ranging from 18% to 51% across regions

5. Feature Engineering

Following the methodology outlined in our project documentation, we created comprehensive engineered features:

Lag Features (Critical for Malaria Prediction):

- Rainfall lags: 1, 2, and 3 months (capturing mosquito breeding cycles)
- Temperature lags: 1, 2, and 3 months (reflecting parasite development periods)
- Humidity lags: 1, 2, and 3 months (accounting for vector survival rates)

- Rationale: Literature consistently shows 1-3 month delays between climate conditions and outbreak manifestation

Rolling Window Features:

- 3-month rolling averages for all climate variables (seasonal smoothing)
- 6-month rolling averages (capturing longer-term climate trends)
- Rolling standard deviations (climate variability indicators)
- Rationale: Malaria transmission responds to sustained climate conditions rather than single-month extremes

Temporal Encoding:

- Cyclical encoding of months using sine/cosine transformation (preserving seasonality)
- Binary seasonal indicators (rainy/dry season classification)
- Time trend variables capturing long-term changes
- Previous outbreak indicators (modeling outbreak clustering effects)

Interaction Features:

- Temperature × Humidity interaction (joint climate suitability)
- Rainfall × Population density interaction (exposure risk amplification)
- Climate suitability index combining all three climate variables weighted by literature-derived coefficients
- Rationale: Malaria transmission depends on complex interactions between multiple factors

Demographic Features:

- Population density categories (Low: <50, Medium: 50-100, High: 100-200, Very High: >200 per km^2)
- Settlement type proxies based on density patterns
- Healthcare access indicators derived from population distribution
- District-specific risk factors encoded numerically

6. Data Transformation

Scaling and Normalization:

- StandardScaler applied to continuous climate variables (ensuring zero mean, unit variance)
- MinMaxScaler for population-based features (preserving interpretability)
- RobustScaler for outlier-prone variables like rainfall extremes
- Rationale: Different algorithms require different scaling approaches for optimal performance

Categorical Encoding:

- One-hot encoding for nominal categories (district, season)
- Ordinal encoding for ranked categories (population density levels)
- Cyclical encoding for temporal features (month, day of year)
- Label encoding for district names (preserving ordinality for tree-based models)

Data Splitting Strategy:

- Temporal split maintaining chronological order: 70% training (2018-2021), 30% testing (2022-2023)
- Stratified sampling within time periods to maintain class distribution
- Spatial cross-validation preparation for model generalization testing
- Rationale: Prevents data leakage while ensuring realistic evaluation of predictive capability

Model Exploration

1. Model Selection

Rationale for Selected Models:

Primary Model: XGBoost (Extreme Gradient Boosting)

- **Strengths:** Superior performance on structured tabular data as demonstrated in Burkina Faso and Gambia studies; built-in handling of missing values; excellent performance on imbalanced datasets through scale_pos_weight; built-in regularization preventing overfitting; feature importance ranking for interpretability
- **Weaknesses:** Requires careful hyperparameter tuning; computationally intensive during training; potential for overfitting without proper validation; less interpretable than simpler models for stakeholders
- **Suitability for Project:** Aligns with literature review findings showing ensemble methods as top performers; handles mixed data types effectively; provides transparency needed for health decision-making

Secondary Model: Random Forest Classifier

- **Strengths:** Robust to overfitting through ensemble averaging; handles missing values naturally; provides clear feature importance rankings; less sensitive to hyperparameters; excellent interpretability for non-technical users
- **Weaknesses:** Can be memory-intensive with large datasets; may not capture complex temporal dependencies as well as gradient boosting; potential bias toward categorical variables with many categories
- **Suitability for Project:** Established performance in malaria prediction literature; suitable for rural deployment with limited computational resources; trusted by healthcare professionals

Baseline Model: Logistic Regression

- **Strengths:** Highly interpretable coefficient-based predictions; fast training and inference; well-understood by medical professionals; minimal computational requirements; provides probability estimates with clear meaning
- **Weaknesses:** Assumes linear relationships between features and log-odds; may underperform on complex non-linear patterns; requires feature scaling; limited ability to capture interactions without explicit feature engineering
- **Suitability for Project:** Serves as interpretable baseline; suitable for settings with limited technical infrastructure; provides clinically meaningful probability estimates

2. Model Training

Hyperparameter Configuration (Based on Literature Best Practices):

XGBoost Configuration: