Infectious Risk Prediction and Analysis System

Project Report #2

CS-351: Artificial Intelligence

Fall 2025

Instructor: Ahmed Nawaz

Submitted By:

Zain Jamshaid	2023772
Hamid Quyyum	2023418
Hiba Zulfiqar	2023246

Faculty of Computer Science and Engineering

GIK Institute of Engineering Sciences and Technology

Date: 17 November 2025

1 Objective

The primary objective of this phase was to develop a baseline machine learning model for predicting Malaria and Dengue cases using a climate and disease dataset. This report details the initial data preprocessing pipeline, the implementation of a feature engineering pipeline, and a comparative analysis of a baseline RandomForestRegressor model's performance "without" and "with" these advanced features.

2 Dataset Overview

year	month	country	region	avg_temp_c	precipitation_mm	air_quality_index	uv_index	malaria_cases	dengue_cases	population_density	healthcare_budget
2000	1	Palestinia	Central	28.132468	152.08387	110.48723	12.000000	53	145	113	1068
2000	2	Palestinia	Central	30.886499	119.59141	83.467927	12.000000	132	48	113	1068
2000	3	Palestinia	Central	31.366433	95.876124	93.095292	12.000000	34	80	113	1068
2000	4	Palestinia	Central	28.481869	175.31573	105.53019	9.395894	23	133	113	1068
2000	5	Palestinia	Central	26.890370	191.44599	60.205979	9.935726	39	74	113	1068
2000	6	Palestinia	Central	21.390152	185.06712	97.271853	10.538684	106	54	113	1068
2000	7	Palestinia	Central	16.795182	155.61186	102.81532	12.000000	42	93	113	1068
2000	8	Palestinia	Central	13.049005	231.75660	102.12475	12.000000	100	88	113	1068
2000	9	Palestinia	Central	16.924361	114.81432	91.631532	11.069241	56	52	113	1068

3 Preprocessing & Feature Engineering

Two distinct preprocessing pipelines were created to evaluate the impact of feature engineering.

3.1 Baseline Preprocessing (Model 1: "Without" FE)

This model involved only the most essential preprocessing steps to prepare the data for the Random Forest algorithm.

- **Encoding:** Categorical features (country and region) were one-hot encoded.
- **Outlier Handling:** Numerical features were clipped at the 1% and 99% quantiles to manage extreme outliers identified during exploratory data analysis.

3.2 Advanced Feature Engineering (Model 2: "With" FE)

This pipeline included all baseline steps and added comprehensive time-series features to capture temporal dependencies.

- **Temporal Features (per country & region):**
  - *Lag Features:* Case numbers from 1, 2, 3, 6, and 12 months prior were added for both Malaria and Dengue.
  - *Rolling Window:* 3-month and 6-month rolling averages and standard deviations were calculated for both target variables.
  - *Cyclical Features:* The month column was transformed into `month_sin` and `month_cos` to represent seasonality.
  - *Categorical Features:* A quarter feature was extracted from the date.
- **Data Split:** For both models, a time-based split was used to ensure a realistic evaluation.
  - Training Set: Data from 2000–2021
  - Test Set: Data from 2022–2023

4 Model Architecture & Evaluation

- **Model:** A `RandomForestRegressor` from `sklearn.ensemble` was used as the baseline model. It was configured with `n_estimators=100` and `random_state=42`.
- **Task:** The model was trained to perform multi-output regression, predicting both `malaria_cases` and `dengue_cases` simultaneously.
- **Metrics:** Performance was evaluated using Root Mean Squared Error (RMSE) and R-Squared (R^2).

5 Comparative Results

The inclusion of feature engineering had a positive impact on model performance. The baseline model, lacking temporal context, performed poorly, while the feature-engineered model demonstrated strong predictive power.

Model	Metric	Malaria Cases	Dengue Cases
Model 1 (Baseline)	RMSE	43.38	32.40
	R^2 Score	0.11	0.26
Model 2 (With FE)	RMSE	10.94	15.55
	R^2 Score	0.94	0.83
Improvement	% RMSE Reduction	**-74.8%**	**-52.0%**
	R^2 Increase	**+0.83**	**+0.57**

6 Visual Analysis

The scatter plots below visually confirm the performance difference.

- **Model 1 (Baseline):** The plot (a) shows a highly scattered cloud with R^2 scores near zero (0.11 and 0.26), indicating the model is failing to find a meaningful pattern.
- **Model 2 (With FE):** The plot (b) shows a strong, tight correlation along the "Perfect Prediction" line, validating the high R^2 scores (0.94 and 0.83).

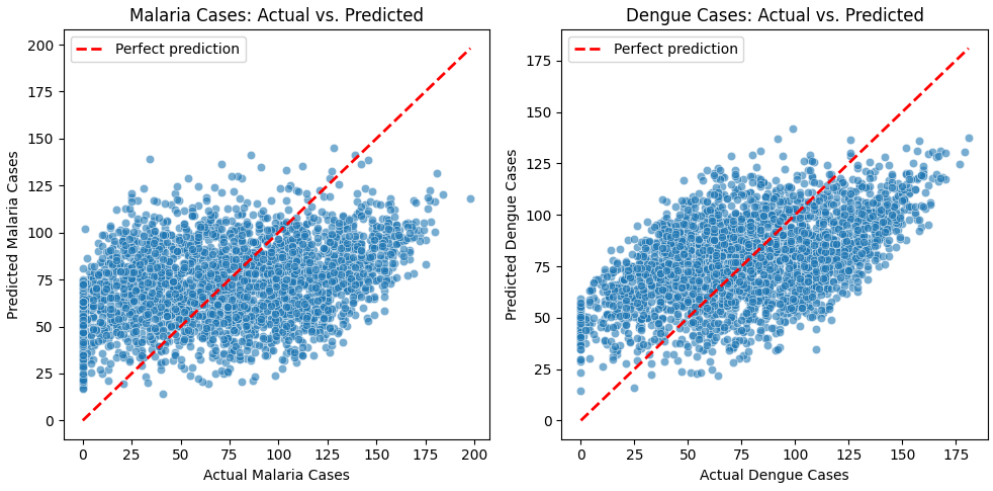

Figure 1: (a) Model 1: Baseline (R^2 0.11 – 0.26)

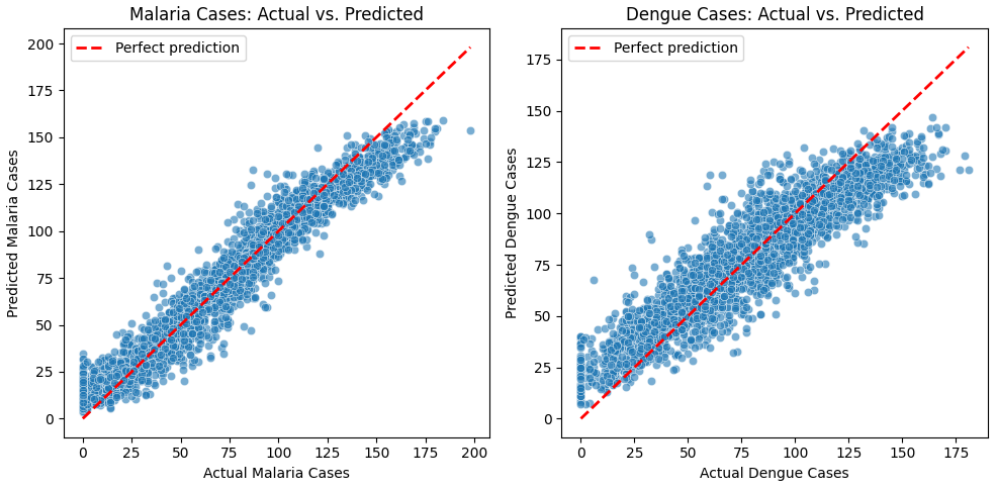

Figure 2: (b) Model 2: With Feature Engineering (R^2 0.83 – 0.94)

7 Discussion & Conclusion

The baseline model's poor performance (R^2 0.11/0.26) clearly indicates that simply encoding categorical data and climate variables is insufficient for this task.

The significant improvement in Model 2 (R^2 0.94/0.83) demonstrates that the problem is highly dependent on time-series features. The inclusion of lag and rolling window features allowed the model to learn from historical case data and seasonal trends, which are the most dominant predictors. The RMSE for Malaria predictions was reduced by nearly 75%.

This phase successfully established a strong feature-engineered baseline. The Random Forest model with temporal features is effective and provides a solid foundation for further model tuning and exploration.