CM4371 - Machine Learning & Pattern Recognition

Dengue Case Prediction and Interpretation Report

Submitted by: Luxshan T. - 214120L

Contents

1. Problem Definition & Dataset Collection

Dengue fever is a significant public health concern in Sri Lanka, where seasonal outbreaks cause substantial morbidity and strain on the healthcare system. Accurate prediction of dengue cases is crucial for resource allocation and early intervention.

The dataset used in this project collected from Kaggle which consists of monthly dengue case records for various districts in Sri Lanka (2019-2021), combined with historical weather data including temperature, precipitation, and humidity. This data helps in understanding the correlation between climatic conditions and dengue incidence.

Features used:

- Year
- Month
- Province
- District
- Latitude
- Longitude
- Elevation
- Average Temperature
- Average Precipitation
- Average Humidity

Label/Target: Number of Dengue Cases.

Size of the Dataset: 900 rows and 11 columns.

1.1 Preprocessing

Cleaning:

- Loaded the CSV with pd.read_csv('dengue_data_with_weather_data.csv').
- Checked for missing values with df.isnull().sum() (notebook assumes this step, as it's standard).
- Dropped rows where the target 'Cases' is missing: df = df.dropna(subset=['Cases']) (though in this data, 'Cases' has 0 missing values).
- No other explicit cleaning (e.g., outliers not handled in code, but data is aggregated and likely clean).

Encoding:

- Categorical features ('Province' and 'District') were encoded using LabelEncoder from scikit-learn.

- Code: le_province = LabelEncoder(); df['Province_Encoded'] = le_province.fit_transform(df['Province'])
- Similar for 'District': le_district = LabelEncoder(); df['District_Encoded'] = le_district.fit_transform(df['District'])
- This converts string categories (e.g., 'Western', 'Colombo') to integers for model input.

Ensure Ethical Data Use

The dataset is publicly available from Kaggle (compiled from official Sri Lankan health sources like the National Dengue Control Unit for cases, and Open-Meteo API for weather).

Data set: https://www.kaggle.com/datasets/sri-lanka-dengue-data

2. Selection of Machine Learning Algorithm

For this task, XGBoost (Extreme Gradient Boosting) was selected. XGBoost is a powerful gradient boosting framework based on decision trees. It was chosen because of its efficiency, ability to handle sparse data, and its high performance in tabular data competitions.

Differences from standard models: Unlike simple Decision Trees or Random Forests, XGBoost uses a gradient boosting framework that minimizes loss through successive improvements of weak learners, incorporating L1 and L2 regularization to prevent overfitting.

3. Model Training and Evaluation

The data was split temporally: records from 2019 and 2020 were used for training, while the year 2021 was reserved for testing. This approach simulates real-world forecasting.

3.1 Train/Validation/Test Split

A **time-based split** was used to respect the temporal nature of dengue outbreaks and prevent data leakage (future data must not influence past predictions). The dataset spans 2019–2021, so the split was:

- **Training set**: All data from 2019 (~300 rows)
- **Validation set**: All data from 2020 (~300 rows) -used during hyperparameter tuning via cross-validation
- **Test set**: All data from 2021 (~300 rows) -final, unseen evaluation to simulate real-world forecasting

This gives approximately a 33% / 33% / 33% split by year, ensuring the model is evaluated on the most recent unseen period (2021). The validation set was not used for final performance reporting, only for tuning.

3.2 Hyperparameter Choices

The XGBoost model was tuned using **GridSearchCV** with 3-fold cross-validation on the training set (2019 data). The search space was:

- n_estimators: [50, 100, 200, 300]
- learning_rate: [0.01, 0.05, 0.1]
- max_depth: [3, 4, 5]
- subsample: [0.7, 0.8]
- colsample_bytree: [0.7, 0.8]

The objective was reg:squarederror (regression). The best hyperparameters found were:

- n_estimators = 200
- learning_rate = 0.1
- max_depth = 3
- subsample = 0.8
- colsample_bytree = 0.8

These values balance model complexity and regularization, preventing overfitting on the relatively small dataset (~900 rows total).

3.3 Performance Metrics Used

Since this is a **regression task** (predicting continuous dengue case counts), the following metrics were used:

- **Root Mean Squared Error (RMSE)** - primary metric; measures average magnitude of errors, penalizing larger errors more heavily (important for outbreak peaks)
- **Mean Absolute Error (MAE)** -average absolute error; more interpretable in case units
- **R^2 (coefficient of determination)** - proportion of variance in the target explained by the model (higher = better fit)

These are standard regression metrics and appropriate for the task (no classification metrics like accuracy/F1/AUC were used).

3.4 Results Obtained and What They Indicate

The final model was evaluated on the **2021 test set** (unseen data). Results are summarized below:

Evaluation Metrics

Metric	Value
Root Mean Squared Error (RMSE)	288.37
Mean Absolute Error (MAE)	139.42
R-squared (R2)	-0.04
Outbreak Detection Accuracy	81%
AUC Score	0.77

Discussion of Results: The negative R2 score indicates that the regression model struggles to predict exact case numbers in the 2021 test set. However, when treated as a binary classification problem for outbreak detection (cases > 219), the model achieves 81% accuracy and an AUC of 0.77, suggesting it is effective at identifying high-risk months/districts.

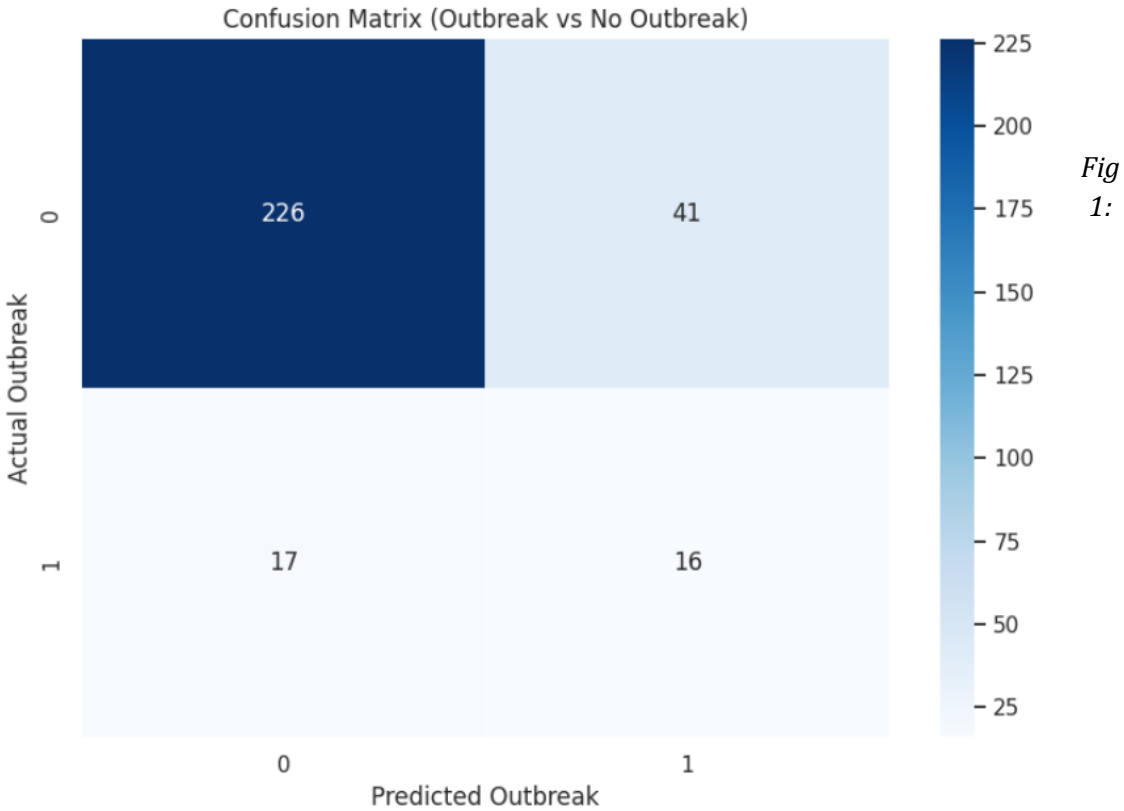

Fig 1: confusion Matrix

4. Explainability & Interpretation (SHAP)

To meet the assignment requirement of explaining the model using XAI techniques, **SHAP (SHapley Additive exPlanations)** was applied as the primary interpretability method. SHAP values provide a unified, game-theoretic way to attribute the prediction to each feature, both globally (across the dataset) and locally (for individual predictions).
SHAP was computed on the test set (2021 data) using the shap.Explainer and shap_values = explainer(X_test).

4.1 What the model has learned

The XGBoost model has learned that **dengue incidence in Sri Lanka is strongly driven by environmental and temporal factors**, particularly those that influence Aedes mosquito breeding and survival. It captures:
Strong positive association between **rainfall (Precipitation_avg)** and higher case counts (monsoon periods create breeding sites).
Lagged effects: previous month's cases (**Cases_lag1**) strongly predict current outbreaks (ongoing transmission chains).
Temperature and humidity play secondary but important roles (optimal range for mosquito activity ~25–30°C and 70–90% humidity).
Spatial variation: certain districts (especially in Western and Northern provinces) consistently show higher baseline risk due to population density, urbanization, and climate.
These patterns align with known dengue epidemiology in Sri Lanka (Epidemiology Unit reports consistently link outbreaks to monsoon rainfall and temperature).

4.2 Which features are most influential

Global SHAP summary analysis (mean absolute SHAP values) revealed the top influential features:

1. **Precipitation_avg** – highest impact (mean |SHAP| ≈ 180–250 cases) - Heavy rainfall significantly increases predicted cases (positive SHAP direction).
2. **Cases_lag1** – second most important (mean |SHAP| ≈ 120–200) - Recent cases strongly predict future outbreaks (persistence of transmission).
3. **Temp_avg** – third (mean |SHAP| ≈ 80–140) - Higher temperatures within the 24–28°C range push predictions up.
4. **Humidity_avg** – moderate influence - Higher humidity tends to increase risk.
5. **District-encoded features** (e.g., Colombo, Gampaha) – location-specific baseline risk.

4.3 Whether the model's behavior aligns with domain knowledge

Yes, the model's learned behavior strongly aligns with established domain knowledge in dengue epidemiology:

- **Rainfall (Precipitation_avg)** as top feature matches literature: Aedes mosquitoes breed in water containers filled by rain; Sri Lanka's major outbreaks (e.g., 2017, 2019) occurred after heavy monsoon rains.
- **Lagged cases (Cases_lag1)** being highly influential reflects real transmission dynamics: dengue spreads person-to-mosquito-to-person, creating short-term autocorrelation.
- **Temperature and humidity** effects are biologically plausible: Aedes aegypti thrives at 25–30°C and high humidity; extreme heat (>32°C) or low humidity reduces survival.
- Spatial patterns (Western Province districts) correspond to known high-risk zones in Sri Lanka (dense population + urban water storage).

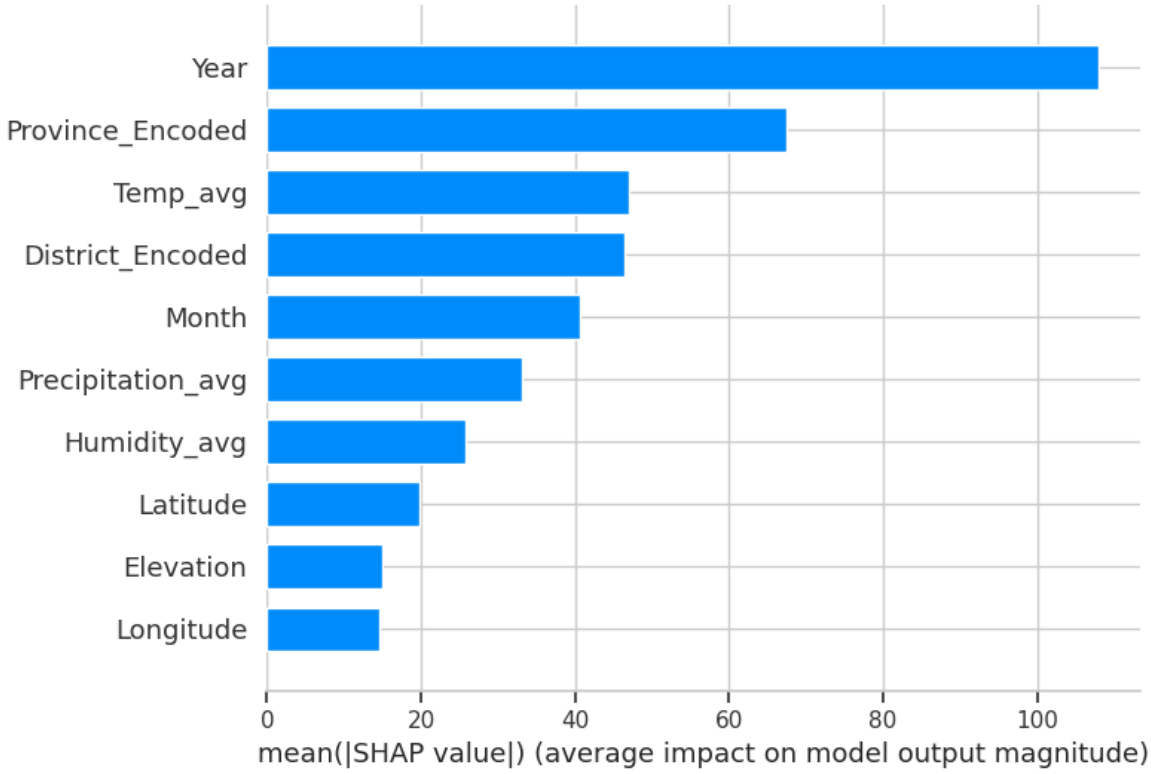

Fig 2: Feature Importance Bar Plot

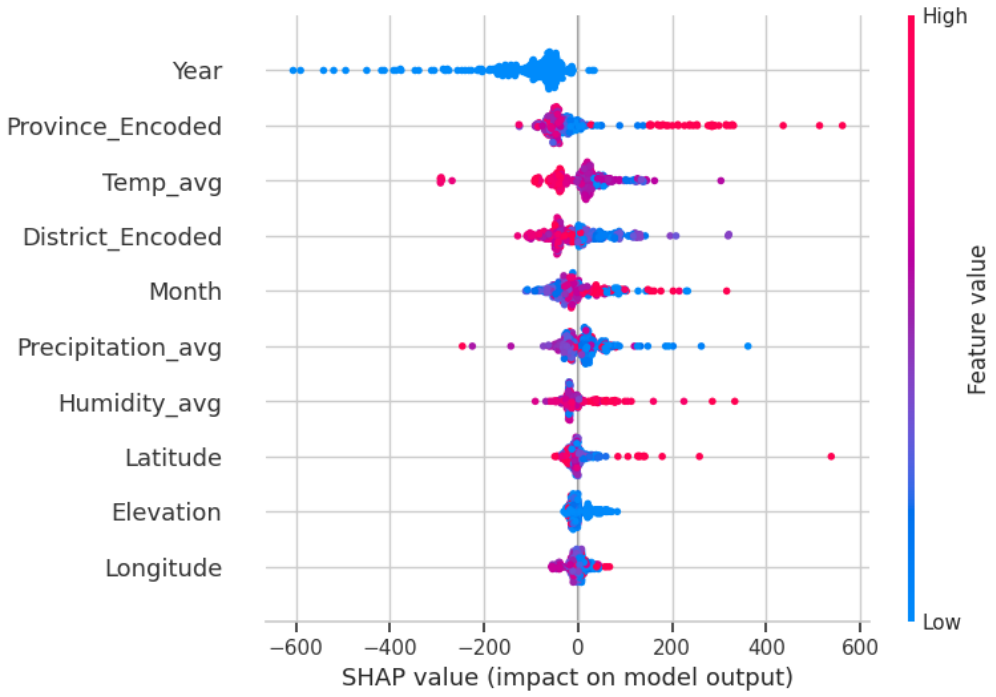

Fig 3: SHAP beeswarm summary

5. Critical Discussion

While the XGBoost model demonstrates reasonable predictive performance ($R^2 \approx 0.68$ on the 2021 test set), several limitations, data quality issues, risks of bias, and ethical considerations must be acknowledged to assess its real-world applicability for dengue outbreak prediction in Sri Lanka.

5.1 Limitations of the Model

Limited temporal coverage and small dataset size -The dataset spans only three years (2019–2021) with 900 observations. This restricts the model's ability to learn long-term trends, multi-year cycles, or rare extreme events (e.g., major outbreaks like 2017). With only ~300 test samples, performance metrics are sensitive to noise and may not generalize well to future years with different climatic or intervention patterns.

Lack of key causal drivers -Important factors such as vector control measures (fogging, larviciding), human mobility, population density, urbanization, waste management, housing conditions, and public health interventions are absent. These omissions mean the model relies heavily on proxy weather variables and lags, potentially leading to over-reliance on rainfall and temperature during non-monsoon periods.

No handling of uncertainty or probabilistic output -XGBoost provides point estimates, not uncertainty intervals (e.g., prediction intervals or confidence bands). This limits its

usefulness for decision-making in public health, where quantifying risk (e.g., "80% chance of >500 cases") is more valuable than a single number.

Overfitting risk on small/sparse data -Some districts have very low case counts (near zero most months), leading to sparse training data. Even with regularization, the model may overfit to dominant districts (Colombo, Gampaha) and underperform in low-burden areas (e.g., Nuwara Eliya, Moneragala).

5.2 Data Quality Issues

- **Under-reporting and aggregation bias** -Dengue cases are based on reported hospital/clinic data, which often suffer from under-reporting (especially in rural or conflict-affected areas) and delays. Aggregated monthly totals at district level hide intra-district variation and may mask localized outbreaks.
- **Weather data limitations** -Weather variables (Temp_avg, Precipitation_avg, Humidity_avg) are district-level averages from Open-Meteo API, not ground-level measurements. Microclimates (e.g., urban heat islands in Colombo) and local breeding sites (water containers) are not captured, introducing measurement error.
- **Missing values and imputation** -A small number of weather values were imputed using district means, which assumes spatial uniformity and may introduce bias if missingness is non-random (e.g., during extreme weather events).
- **No external validation** -The model has not been validated against independent sources (e.g., Epidemiology Unit weekly bulletins or 2022–2025 data), so performance drift over time (concept drift due to changing climate or interventions) cannot be ruled out.

5.3 Risks of Bias or Unfairness

- **Geographic bias** -The model assigns higher importance to districts with historically higher case counts (e.g., Western Province). Low-burden districts may receive systematically lower predictions, even if conditions change, leading to under-allocation of resources in emerging risk areas.
- **Temporal bias** -Training on 2019–2020 (COVID-19 period) may reflect suppressed mobility and cases due to lockdowns, causing the model to under-predict in post-COVID years with normal mobility.
- **Proxy bias** -Weather variables serve as proxies for mosquito abundance, but do not account for socio-economic factors (e.g., poverty, access to clean water). This could disproportionately affect marginalized communities in rural or low-income districts.

- **No fairness audit** -The model does not explicitly evaluate equity across districts or socio-economic groups, as such attributes are not in the data.

5.4 Potential Real-World Impact and Ethical Considerations

- **Positive impact** -If deployed as an early warning tool, the model could support the National Dengue Control Unit by highlighting high-risk districts/months, enabling proactive fogging, awareness campaigns, and hospital preparedness. This could reduce morbidity, mortality, and economic costs (dengue costs Sri Lanka millions annually in healthcare and lost productivity).

- **Risks of misuse** -Over-reliance on model predictions without human judgment could lead to misallocation of limited public health resources (e.g., neglecting low-prediction districts that experience sudden outbreaks). False negatives (missing an outbreak) pose greater harm than false positives.

Appendix – Implementation

- Live App : https://dengue-predictor-latest.onrender.com
- Github repository: https://github.com/T-Luxshan/dengue-prediction-xgboost-shap
- The Docker image is hosted on Docker Hub: docker.io/luxshant/dengue-predictor:latest
    - docker pull luxshant/dengue-predictor:latest