Dengue Fever Weekly Risk Alert (Thailand)

Team: *Group 8*
Course: Capstone FTL — Idea Proposal Submission

1) Project Idea

We will build a small system that gives weekly dengue risk alerts for each district/province in Thailand.
The tool looks at:

- **Weather**: rain, temperature, humidity, wet-spell days.
- **Behavior signals**: Google searches for “dengue,” “ไข้เลือดออก,” “ยุงลาย,” etc.
- **Past dengue cases** to check accuracy.

It outputs:

- A risk score (0–1) and class (Low / Medium / High) per area/week.
- A map + table with the Top-K high-risk districts.
- A short why-this-area note using simple explanations (top 3 drivers from SHAP).

Goal: help local teams act earlier (larval control, community notice, school clean-ups) with a clear, easy-to-read dashboard.

2) Relevance to SDGs

- **SDG 3 – Good Health & Well-Being:** early warning can cut cases and hospital load.
- **SDG 11 – Sustainable Cities & Communities:** supports city planning for vector control.
- **SDG 13 – Climate Action:** links climate patterns (rain/heat) to disease risk and helps climate-health adaptation.

3) Literature Examples (brief)

- Weather + Search data for dengue forecasting: Studies show that combining rain/temperature with Google Trends improves short-term dengue prediction at city level; tree-based models often beat simple baselines.

- Thailand/SEA climate-driven dengue models: Prior work uses rainfall, temp, humidity, ENSO to forecast dengue 1–8 weeks ahead; adding human-behavior signals further improves timeliness.

4) Describe Your Data

Sources (open):

- **ERA5 / GPM (satellite) / TRMM**: daily rain, temp, humidity → aggregated to **weekly by district**.

- **Google Trends**: weekly relative search volume for dengue-related keywords (Thai + English).

- **Historical dengue cases**: provincial/district time series from Thai MOPH/WHO (where public).

Format & size: CSV/Parquet tables with columns: `date_week, province/district_id, rain_mm, temp_C, humidity_%, wetspell_days, search_dengue, …, cases` (if available). About **3–10 years** of weekly rows × **~77 provinces** (or districts when possible).

Pre-processing:

- Align to **ISO week**; fill small gaps; create **lags** (e.g., rain at t-1..t-4), **rolling means/sums**, **wet-spell counts**, and **seasonality features** (week-of-year, holidays, ENSO index if used).

- Normalise features; train/validation **time-split** (walk-forward).

5) Approach (Machine Learning)

Why ML (not deep learning first): our data are **tabular + short time series**; **tree ensembles** are strong, fast, and **interpretable**.

Baseline pipeline:

1. **Models:** Logistic Regression (simple), **XGBoost/LightGBM** (main), Random Forest (backup).
2. **Target:** High-risk label per area/week (e.g., top 20% of historical cases or exceed a moving threshold).
3. **Training:** walk-forward cross-validation by week; province-aware splits to avoid leakage.
4. **Explainability: SHAP** to show top drivers each week/area.
5. **Evaluation:** AUROC/PR-AUC, F1 for High-risk, **Brier score**, and a **Top-K hit rate** ("Did we flag the districts that later spiked?").
6. **Output:** a **weekly map and table** plus a short reason line (e.g., "High rain last 2 weeks + high humidity + search spike").

Stretch goals:

- Add **LSTM/Temporal Fusion Transformer** if longer series become available.
- Add **entomology proxies** (NDVI, standing-water indices) from satellite images.
- Simple **SMS/LINE** alert for local partners.

Deliverables (MVP):

- One-page **dashboard** (map + Top-K list + reasons).
- Clean **CSV features** and code to re-run **weekly**.
- Short **user guide** for public-health staff.