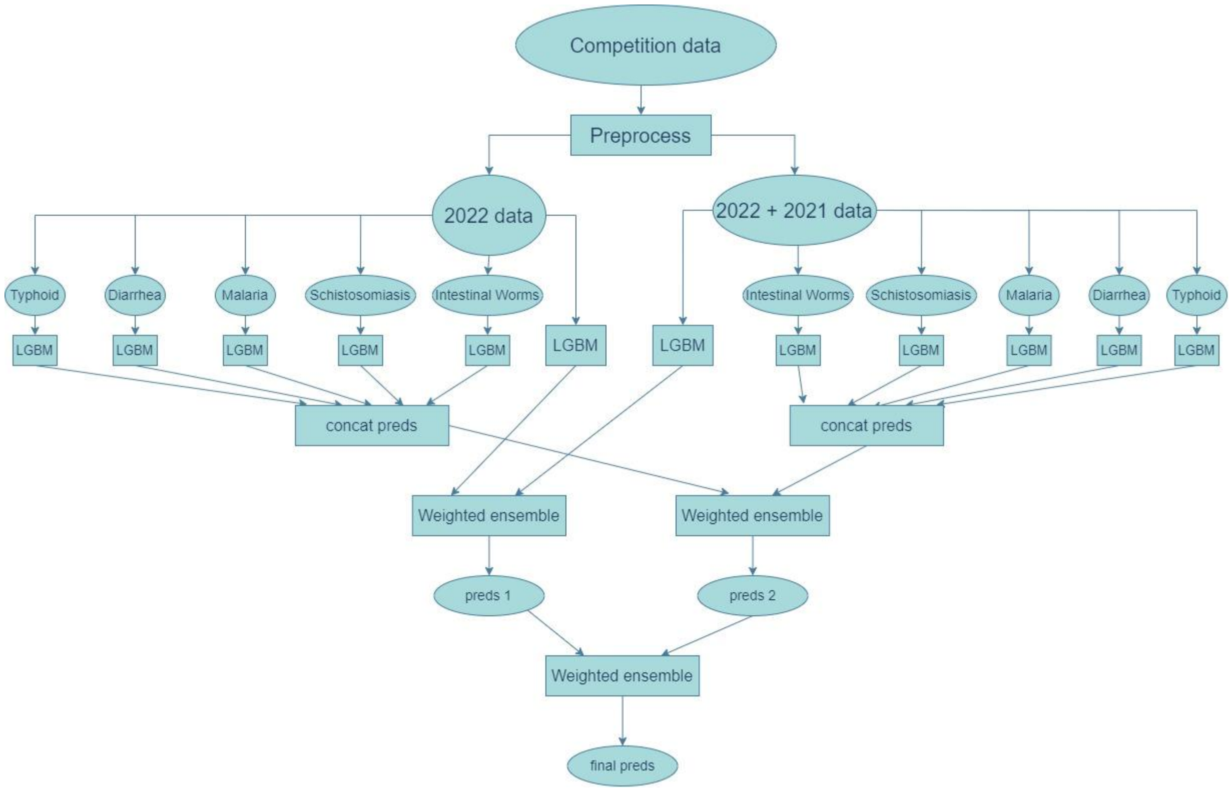
SUA Outsmarting Outbreaks Challenge – 2nd Place solution

1. Overview

The goal of this challenge was to develop a machine learning model that can predict the outbreaks of climate-sensitive waterborne diseases. A two-stage pipeline was used; in stage 1 we trained multiple Lightgbm models on parts of the training set divided by the year of the outbreak and the type of disease. In stage 2 we added the pseudo labels (predictions) of the test that were generated from stage 1 to the training set and then repeated the same training of stage 1.

2. Architecture diagram



3. ETL process

- **Data cleaning**

There are multiple samples that have the same features per id for each year, but the target is different sometimes, it is advisable in this case to not train a model with the same input and different outputs, so we will also make the output unique and select the mean or max target per id.

- **Target reduction**

After some analysis, we found out that the mean of the target per year has a constant bias:

Year	Mean of target
2019	18.295832
2020	15.540424
2021	13.049686
2022	10.775961

So, we will also reduce the target values by 3 for the next year 2023 to respect the same bias.

- **Feature engineering**

Multiple features were generated from the numerical columns and added:

- ✓ PCA components
- ✓ K-means clustering
- ✓ Random tree embeddings

- **Data normalization**

Row-wise scaling was used with norm l2.

- **Cross validation scheme**

5 StratifiedKFold, the stratification was performed on location + disease so we will have the same proportion of those two in each fold. It was computed only on the last year (2022), old historical data for year 2019 and 2020 were not used at all because they made this cv worse, only 2021 was used, some models were trained with only 2022 data and others with 2022 + 2021 and ensembled with those of 2022 to add more diversification.

- **Postprocessing**

Predictions were rounded to an integer, it improved cv and also worked well on lb.

4. Data modeling

The *Lightgbm* model was trained with early stopping.

Training Parameters for all models

- objective: mae
- metric: mae
- learning_rate: 0.03
- max_depth: 5
- seed: 42

- boosting: goss
- top_rate: 0.3

5. Inference

All training and inference were done in the same notebook.

6. Run time

Train and inference time : less than half an hour

7. Performance metrics

- **Public leaderboard :** 5.391666666
- **Private leaderboard :** 6.491075981

8. Remarks

- There is not much difference between the submission generated from stage 1 and the one from stage 2

  Stage 1: Public LB: 5.398, Private LB: 6.472
  Stage 2: Public LB: 5.392, Private LB: 6.491

- Location was not used as a feature because there is new location in the test set that doesn't exist in the training set.