Title: Predicting Areas with High or Low Typhoid Prevalence Using Machine Learning

# Introduction

Typhoid fever remains a significant public health concern in many regions worldwide. Accurately identifying areas with high or low typhoid prevalence is crucial for implementing targeted interventions and allocating resources effectively. In this project, we aimed to develop a machine learning model to predict areas with high or low typhoid prevalence based on district-level data from a krigged dataset spanning 2017.

Our approach involved utilizing features derived from the average number of typhoid cases in each district, along with additional environmental and demographic factors.

# Data Exploration and Feature Engineering

The features in the dataset are as follows:

* **FID\_**: Feature ID or unique identifier for each district.
* **DNAME\_2011**: District name.
* **Typh\_Inc**: Number of typhoid incidences in the district (target variable).
* **X\_coord**, **Y\_coord**: Geographic coordinates (longitude and latitude) of the district.
* **HH\_Wash**: Percentage of population exercising hand washing practice after using the toilet.
* **PH\_Lands**: Percentage of population that stays in highlands.
* **P\_Density**: Population density in the district.
* **Urban\_level**: Proportion of urbanization in the district.
* **ARainfall**: Average annual rainfall in the district.
* **Temp\_Max**, **Temp\_Min**: Maximum and minimum temperatures in the district.
* **Typh\_Rate**: Typhoid incidence rate per capita.
* **Pn\_Floods**: Proportion of people affected by floods per district.
* **P\_male**, **P\_Female**: Male and female population counts in the district.
* **Typh\_Per**
* **OBJECTID**: Object ID or unique identifier for each record.

Using the **Typh\_Inc** variable, we decided to create a feature called **Typh\_Inc\_Label** which we eventually created a machine learning model for to try and predict. The **Typh\_Inc\_Label** is a feature that assigns a value of high, to indicate high typhoid prevalence, or low, to indicate the opposite.

We conducted exploratory data analysis by leveraging the folium library to plot maps showing the distribution of typhoid cases across districts, as well as the different areas with high or low prevalence. This was done to get insights and identify potential predictors of typhoid prevalence.

Feature engineering was also performed to create new features such as temperature range, population, and interaction terms between temperature and population density. We discussed as a group which features to use as the dependent variables, and of course, used the **Typh\_Inc\_Label** as the dependent variable. Features we deemed unnecessary were dropped. These included **'FID\_', 'DNAME\_2011', 'X\_coord', 'Y\_coord', 'OBJECTID', 'Typh\_Inc', 'Population', 'Typh\_Rate', 'P\_male', 'P\_Female'.**

Encoding was carried out on the dependent variable so that we could pass it into the models for training.

# Model Building and Evaluation

## Random Forest Classifier

We selected the Random Forest Classifier as our primary model due to its ability to handle complex datasets and provide robust predictions.

Hyperparameter tuning was conducted using GridSearchCV to optimize the model's performance.

The model was trained and evaluated using 10-fold cross-validation to ensure reliable estimates of performance.

Evaluation metrics including accuracy, precision, recall, and F1 score were calculated to assess the model's ability to predict areas with high or low typhoid prevalence.

## XGBoost Classifier

In addition to the Random Forest Classifier, we experimented with an XGBoost Classifier to compare performance.

Similar hyperparameter tuning and evaluation procedures were applied to the XGBoost model

# Results

The XGBoost Classifier demonstrated superior performance compared to the Random Forest Classifier, achieving higher accuracy, precision, recall, and F1 score.

Key predictors of typhoid prevalence identified by the XGBoost model included HH\_Wash, PH\_Lands, P\_Density, Urban\_leve, ARainfall and Temp\_Max.

The model successfully differentiated between areas with high and low typhoid prevalence, providing valuable insights for public health planning and intervention strategies.

# Conclusion

In conclusion, our machine learning approach effectively predicted areas with high or low typhoid prevalence using district-level data from the krigged dataset. By leveraging features derived from average typhoid cases per district along with environmental and demographic factors, the XGBoost Classifier provided actionable insights for public health authorities. These findings can inform targeted interventions aimed at reducing typhoid burden and improving public health outcomes in at-risk regions.