679 Final Report

Beiming Yu

Yangyu Chen

1. Introduction:

Dengue fever is a viral disease transmitted by Aedes mosquitoes and poses a growing threat to public health in tropical and subtropical regions. Its outbreaks are often driven by complex interactions between environmental, climatic, and ecological factors like the temperature humidity, since the activity of mosquitoes is influenced by those factors. Accurate prediction of dengue case counts might play a crucial role in resource allocation, public health planning, and outbreak prevention. This project leverages a dataset of weekly dengue case counts and meteorological features from two Latin American cities—San Juan, Puerto Rico and Iquitos, Peru—to build predictive models using machine learning. We had weekly records for the two cities over the following time periods:

San Juan (sj): from week 18 of 1990 to week 24 of 2008

Iquitos (iq): from week 44 of 2000 to week 25 of 2010

These time frames cover multiple seasonal cycles and dengue outbreak periods, providing a rich temporal structure for modeling the disease cases.

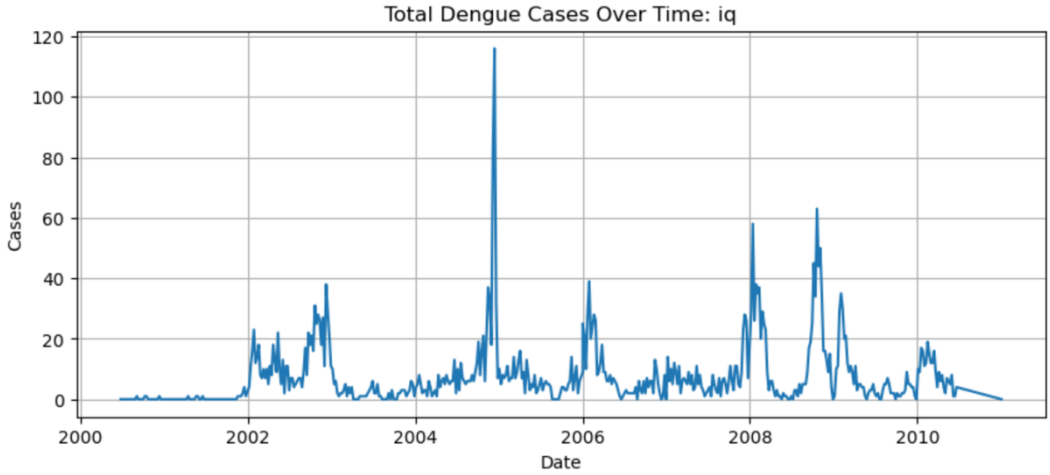
By analyzing patterns in temperature, humidity, precipitation, and vegetation indices, and incorporating temporal dynamics through lag features and rolling forecasts, we tried to identify reliable modeling strategies to forecast dengue outbreaks. The work explores a range of statistical and machine learning methods—including Lasso, Ridge regression, Random Forest, and XGBoost—with the goal of minimizing prediction error and improving real-world applicability through techniques such as overfitting mitigation and post-prediction adjustment. Although we got a great model with small MAE, and the difference of predictions between training set and testing set is small. It did not work well in the competition.

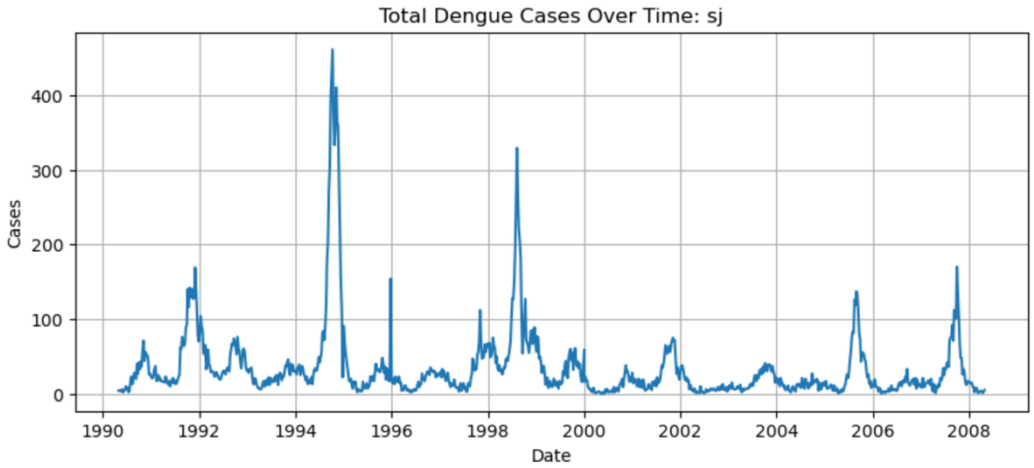
2.EDA and Data cleaning:

Before modeling, our first step is always doing exploratory data analysis (EDA) and data cleaning to ensure the dataset was reliable and well-structured. Missing values were addressed by forward-filling within each city to maintain the temporal consistency of the features. However, if a feature contains more than 30% missing values, it will be dropped to reduce noise and prevent bias. We removed the week_start_date column.

Since the dataset had information from the two cities, models were built independently for the two cities. As a result, the city characteristic was neither one-hot encoded nor used as a predictor variable.

We then visualized dengue case trends over time for each city, revealing seasonal patterns and periodic surges in case counts. We draw a plot that shows the relationship between cases and time, since features like temperature and humidity are influenced by the time. We found two cities show different patterns, so we may train a separate model for each of them.





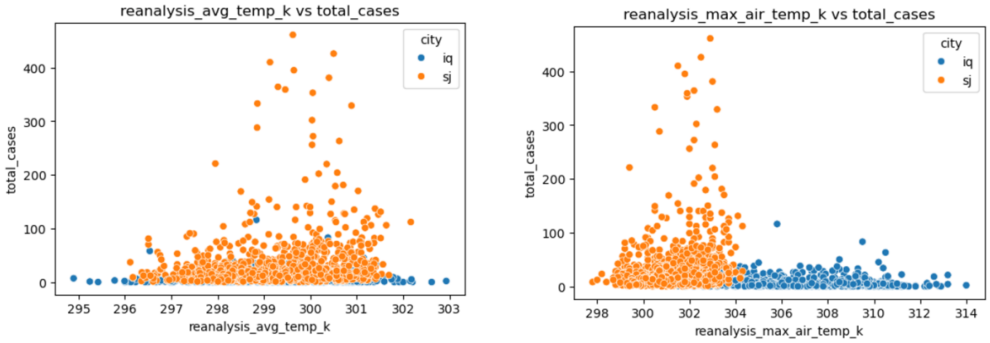
We also drew scatter plots about the relationship between temperature and dengue case counts in the two cities.

In the left plot, we observe the relationship between average air temperature in Kelvin and total_cases. In San Juan (orange dots), dengue outbreaks tend to occur when average temperatures are in the range of 298–300 K (approximately 25–27°C). The case counts rise significantly within this band, which shows a possible optimal temperature range for mosquito activity and virus transmission. In contrast, Iquitos (blue dots) shows

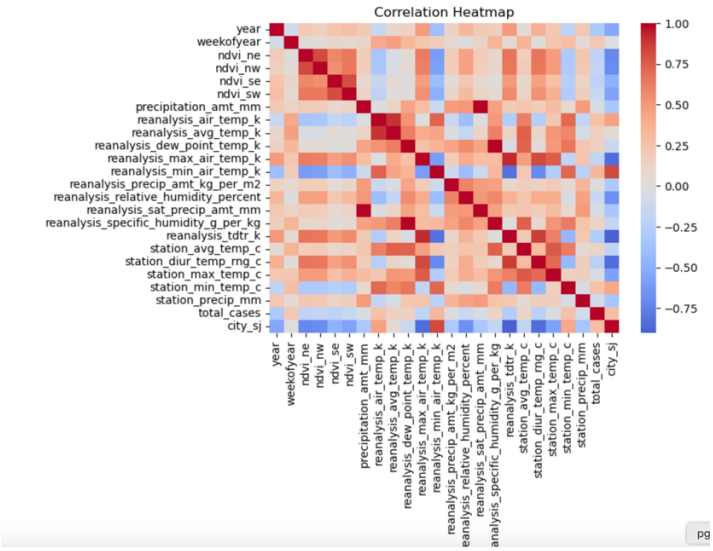
much lower variation in temperature and fewer extreme dengue outbreaks, with most case counts remaining low regardless of temperature.

The right plot shows maximum air temperature versus total_cases. A similar pattern emerges for San Juan, where high case counts cluster around 302–304 K (~29–31°C). Again, we see a spike in cases in this narrow temperature range, reinforcing the hypothesis that dengue transmission is sensitive to specific thermal conditions. For Iquitos, there is no clear peak, and the spread is more uniform with lower case counts overall.

These plots suggest that temperature plays a more prominent role in driving dengue outbreaks in San Juan than in Iquitos, and that outbreaks tend to intensify within specific temperature thresholds. This insight can help inform feature selection and targeted forecasting strategies.



Correlation heatmaps were used to examine relationships among climate variables and identify redundant features. These visual insights informed the feature selection process and helped us drop highly correlated variables, simplifying model input without sacrificing information.



The following features were removed due to high multicollinearity: reanalysis_avg_temp_k, reanalysis_sat_precip_amt_mm, reanalysis_specific_humidity_g_per_kg, reanalysis_tdtr_k, and city_sj. These variables are strongly correlated with other predictors in the dataset, which can cause issues such as inflated standard errors and unstable coefficient estimates in regression models. For example, reanalysis_avg_temp_k is likely to change with station_avg_temp_c, while reanalysis_specific_humidity_g_per_kg and reanalysis_tdtr_k are often collinear with dew point or temperature-derived variables. city_sj was removed as we have already mentioned at the beginning of this section. Removing these features helps improve model robustness and interpretability.

3 Model Sleection and Evaluation:

3.1 Transforming the dataset:

To model the number of dengue cases, we began by thoroughly cleaning and transforming the dataset to ensure its suitability for predictive modeling. After merging the training features with the target labels, we sorted the dataset by city, year, and week of year, and applied forward-filling within each city to impute missing values in a temporally consistent way. Because of the data set shows a clear seasonality which means the number of cases heavily relies on the cases in previous week, we want to make the data chronologically for potential time series analyzing. The target variable, total_cases, was log-transformed using log1p to reduce skewness and stabilize variance, since there is one or two burst of cases in both cities.

3.2 Initial Model Selection:

For initial model selection, we explored a set of baseline and regularized models to identify a robust starting point. We start with a linear model that was using as a baselines model. A LASSO regression was chosen for its ability to perform variable selection and prevent overfitting in the presence of many potentially correlated predictors. The penalty allows us to identify a compact set of influential features and serves as a useful benchmark for more complex models.

We also consider the non-linear relationship between the predictors and target variables. Thus, Random Forest was included as a non-parametric, tree-based ensemble model that can automatically capture nonlinear relationships and interactions between variables without extensive preprocessing. It is also relatively robust to noise and outliers. In parallel, we tested XGBoost, a powerful gradient boosting framework that often yields strong performance in tabular prediction tasks. XGBoost provides advanced regularization, handles missing values internally, and allows fine-tuning of model complexity through hyperparameters like learning rate, tree depth, and number

of estimators.

These three models were selected to provide a diverse view of model performance—LASSO offering a linear and interpretable baseline, Random Forest providing a more flexible, ensemble-based approach, and XGBoost representing a state-of-the-art method in supervised learning. By evaluating these models in parallel, we were able to conclude a direction to go further.

```
LASSO MAE: 19.98, RMSE: 35.46
Random Forest MAE: 11.83, RMSE: 24.00
XGBoost MAE: 10.95, RMSE: 21.38
GLMM MAE: 17.79, RMSE: 33.07
```

Based on the evaluation metrics shown, the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), XGBoost outperformed the other models, achieving the lowest MAE (10.95) and RMSE (21.38), indicating more accurate and stable predictions. Random Forest followed closely with strong performance (MAE: 11.83, RMSE: 24.00). In contrast, LASSO regression showed considerably higher error (MAE: 19.98, RMSE: 35.46), likely due to its linear structure and limitations in handling complex feature interactions.

3.3 Overfitting

However, the metric of the models are not accurate at this time, since we split the into 2 parts, 70% of the data are used as training set and the remaining 30% is used as testing data. We than checked the performance of model in both sets.

```
LASSO Performance:
Train MAE: 21.06, Test MAE: 19.93
Train RMSE: 40.86, Test RMSE: 35.52

Random Forest Performance:
Train MAE: 4.36, Test MAE: 11.71
Train RMSE: 7.88, Test RMSE: 23.95

XGBoost Performance:
Train MAE: 0.22, Test MAE: 10.89
Train RMSE: 0.32, Test RMSE: 21.75
```

The clear difference of MAE in training set and testing set indicate a overfitting issue might happen in the tree models like random forest and XGboost model. Then I tried to change the parameter like the tree depth of the model. Using XGboost as an example, here is result of XGboost after regularization

```
XGBoost (Regularized):
Train MAE: 9.22, Test MAE: 12.08
Train RMSE: 16.00, Test RMSE: 25.31
```

The difference between MAE in training data and test data reduced to a acceptable range at that time. Then we move to the next step that trying to increase the performance of the model.

3.4 Lagging

Adding lag features in time-series modeling, especially for predicting disease spread like dengue, is a strategic step to incorporate temporal dependency — the idea that the number of cases today is often influenced by what happened in the recent past. In the dataset provided, we are given weekly observations of dengue cases (total_cases) for two cities along with environment change. However, the virus transmission process does not respond instantly to climate changes or infection surges. For example, an increase in temperature or rainfall may lead to a surge in mosquito breeding, but the impact on human infection typically appears with a delay — due to incubation periods and the time it takes for vectors and hosts to interact. Similarly, a spike in cases one week could indicate a growing outbreak that is likely to continue into subsequent weeks unless interventions are applied.

By introducing lagged features — such as lag1_cases, lag2_cases, lag3_cases — we allow the model to learn from recent historical trends in dengue case counts. This enhances the model's ability to recognize outbreak dynamics and improves forecasting accuracy, especially for short-term predictions. Lags help capture inertia in the system and feedback effects that wouldn't be apparent from weather and environmental features alone.

	Model	Train MAE	Test MAE	Train RMSE	Test RMSE
0	Random Forest (Lag1-3)	6.680484	10.544707	15.158454	14.178725
1	Ridge Regression (Lag1-3)	9.754908	8.038412	19.711770	11.840746
2	XGBoost (Lag1-3)	7.302730	7.549099	14.175822	12.519561

The MAE and RMSE are stable in both training set and testing set.

3.5 Deep Leaning Models:

We also tried to apply deep learning models in this project aimed to capture complex relationships in dengue case progression, particularly using LSTM and GRU architectures. The independent LSTM model for San Juan, leveraging a 12-week input window, demonstrated strong performance (MAE: 15.63, RMSE: 25.88), validating the suitability of deep recurrent structures for long time-series with richer historical data. To enhance the result, a function-based LSTM framework was implemented to separately model both San Juan and Iquitos, adjusting window lengths based on data availability.

There is a surprising result produced, while the San Juan model maintained consistent performance (MAE: 15.96), the Iquitos model achieved a notably low MAE of 8.45, though it suffered from a higher MAPE due to the city's lower-case counts.

Additionally, a GRU-based model was explored as an alternative. Although GRU performed comparably in Iquitos (MAE: 8.20), its results for San Juan showed signs of underfitting (MAE: 19.94), suggesting that GRUs may be less effective in capturing long-term dependencies in richer datasets. Overall, deep learning models, especially LSTMs, demonstrated promising predictive capacity, particularly for cities with sufficient historical depth, though they require careful tuning and data preprocessing to balance accuracy and generalization.

However, the size of the dataset is not large enough for more complicated models and there is a strong signal for overfitting happened in the result for these models. The outcome of the competition with the models are not great as well.

4. Results and other improvement:

When we initially applied predictive models such as Random Forest and XGBoost to the dengue dataset, we observed that the resulting Mean Absolute Error (MAE) on the test data remained relatively high which is consistently close to 30. This was disappointing, especially considering that this error level was very similar to what we observed when fitting the same models directly to the training set without any adjustments or safeguards against overfitting. In essence, the models were not generalizing well and were likely memorizing patterns in the training data that did not translate effectively to unseen data.

There are several potential reasons for this poor generalization. First, the nature of the dengue dataset poses a significant challenge: the disease dynamics are influenced by a complex interplay of environmental factors, time-lagged effects, and possibly unobserved confounders such as public health interventions or social behaviors. Moreover, the data available is relatively small for high-capacity models like Random Forests or XGBoost, which can easily overfit if not carefully regularized. The lag features we introduced may have helped slightly, but they may not have been sufficient to stabilize predictions when environmental variables fluctuate sharply across seasons or years. Also, we noticed that there might be a robust of the disease in some time spot, that the number of patients increased significantly in short time, which extend MAE greatly.

Because these challenges, we shifted our focus to exploring simpler and more stable models. This included trying regularized linear models like LASSO and Ridge regression, as well as applying dimensionality reduction techniques such as PCA. While these

methods provided more conservative and interpretable results, they did not significantly outperform the ensemble models in raw error metrics. However, they helped us better understand the relationship between variables and confirmed that some models were indeed overfitting the noise.

```
🔷 PCA + Random Forest
MAE (RF + PCA): 36.14

🔷 Ridge Regression
MAE (Ridge): 35.57

🔷 LASSO Regression
MAE (LASSO): 32.99
```

Ultimately, we achieved our best performance using LightGBM, a gradient boosting framework that balances speed, regularization, and robustness. LightGBM handled the structure of the data well, especially when coupled with careful hyperparameter tuning and lag features. It produced lower MAE and RMSE values than other models we tried, while also being more computationally efficient. Although the result is still not perfect that the MAE is 27, it is a great progress for us.

Conclusion:

In this project, we explored a range of models to predict dengue cases, starting with extensive data cleaning, feature engineering, and lag creation to capture temporal patterns. Classical models like Random Forest, Ridge, and XGBoost initially suffered from overfitting, but performance improved through hyperparameter tuning and regularization. Surprsingly, LightGBM ultimately delivered the best classical performance.

We also implemented deep learning models using LSTM and GRU architectures. These models, especially LSTMs, performed well in cities with longer historical data like San Juan, effectively capturing sequential trends. From this work, we learned that model performance depends heavily on data characteristics and thoughtful preprocessing. Future improvements could include better model in predicting time series and maybe a carefully constructed GLMM model with splines.