DengAI: Predicting Weekly Dengue Cases with Polynomial Regression, XGBoost, and LSTM

BLG527E Machine Learning Term Project (Fall 2025–2026)

Ömer Malik Kalembaşı
704241038
Istanbul Technical University
kalembasi18@itu.edu.tr

Tanalp Oğuz
708241019
Istanbul Technical University
oguz25@itu.edu.tr

Abstract

This study focuses on weekly dengue case forecasting using the DrivenData DengAI[1] benchmark, following the requirements of the term project. As a baseline, a high-degree polynomial regression model is implemented to examine the limits of explicit non-linear modeling on time-series data. In addition, several machine learning approaches are evaluated, including Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, XGBoost regression, and a heterogeneous ensemble combining these models.

All models are trained separately for San Juan and Iquitos using a consistent feature engineering pipeline incorporating seasonal encodings, lagged variables, rolling statistics, and imbalance-aware sampling. Performance is assessed using Mean Absolute Error (MAE) on city-specific validation splits and on the official competition leaderboard. Among the evaluated methods, a tuned XGBoost model achieves the best performance with a leaderboard MAE of **22.62** and a rank of **574**.

As required by the project, a second model formulation is implemented using the XGBoost framework to predict dengue cases several weeks in advance based solely on information available at the current time. The ahead-of-time analysis identifies the optimal prediction horizon as $t+12$ weeks for San Juan and $t+2$ weeks for Iquitos, resulting in a final leaderboard MAE of **28.44** and illustrating the trade-off between early prediction and forecasting accuracy.

Keywords

Dengue forecasting, time series, k-polynomial regression, XGBoost, LSTM, SVM, ensemble of models, DrivenData DengAI

1 Introduction

Dengue fever is a rapidly spreading mosquito-borne disease that poses a major public health challenge, particularly in tropical and subtropical regions. Its transmission dynamics are influenced by complex interactions between climatic, environmental, and temporal factors, making reliable forecasting a non-trivial task. Accurate short-term and early warning predictions can support proactive public health interventions by enabling timely resource allocation and outbreak preparedness.

Recent advances in machine learning have shown strong potential for modeling infectious disease dynamics, especially in the presence of non-linear relationships and high-dimensional environmental data. Tree-based ensemble methods and recurrent neural networks have been successfully applied to dengue forecasting, while simpler parametric models remain valuable as transparent baselines for comparison.

In this term project, we study the DrivenData DengAI benchmark, which involves predicting weekly dengue cases for two epidemiologically distinct cities: San Juan (Puerto Rico) and Iquitos (Peru). Due to their differing climatic patterns and transmission behaviors, all models are trained and evaluated separately for each city. Following the project requirements, we first implement a high-degree polynomial regression model as a baseline. We then explore more expressive approaches, including Support Vector Regression (SVR), Long Short-Term Memory (LSTM) networks, XGBoost regression, and a heterogeneous ensemble combining these models.

Beyond standard one-week-ahead prediction, we also investigate the problem of forecasting dengue cases several weeks in advance. By evaluating multiple prediction horizons, we aim to identify the most effective look-ahead period for each city. Through systematic comparison of models, feature engineering strategies, and forecasting horizons, this work provides insights into the strengths and limitations of different learning-based approaches for dengue case prediction.

2 Dataset and Problem Definition

2.1 Dataset

In our case, the data is taken from the DengAI dataset provided by DrivenData, which consists of epidemiologic and environmental observations for two geographically and climatologically diverse cities: San Juan (Puerto Rico) and Iquitos (Peru). The dataset includes: meteorological parameters, vegetation parameters, and climate-related variables. Such variables include temperature, rainfall, humidity, and vegetation indices. The target variable to be predicted is the total number of reported dengue cases per week for each city.

Due to the large differences in climate patterns and dengue dynamics between the two locations, the dataset is treated as two separate time series, and all models are trained independently for each city.

2.2 Evaluation Metric

Model evaluation is performed using Mean Absolute Error (MAE), which is the official performance metric of the DengAI competition. We report MAE scores separately for San Juan and Iquitos, and where applicable, we also evaluate the combined performance across both cities.

,
2026.

3 Experimental Analysis of Predictive Models

In this section, we evaluate four distinct modeling strategies ranging from required baselines to deep learning approaches. We assess the performance of each model based on Mean Absolute Error (MAE) using city-specific validation splits.

3.1 k-Polynomial Regression

As a required baseline, we employ a polynomial regression model to approximate non-linear relationships between environmental features and weekly dengue case counts.

Model formulation. Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the feature vector at week t, and y_t the corresponding number of dengue cases. The model learns a polynomial function of degree k:

$$\hat{y}_t = f_k(\mathbf{x}_t) = \mathbf{w}^\top \phi_k(\mathbf{x}_t),$$

where $\phi_k(\cdot)$ denotes the polynomial feature expansion up to degree k, and $\mathbf{w}$ is the weight vector. Ridge regularization is applied to control variance and reduce overfitting introduced by polynomial expansion.

The polynomial regression process is implemented independently for each cities San Juan and Iquitos, following described workflow below. First, none predictive variables such as `city` and `week_start_date` are extracted from the feature set. To model seasonality, cyclical week features are introduced using sine and cosine transformations of the week-of-year variable. Temporal dependency is captured through the generation of lag features and rolling mean statistics derived from historical dengue case counts. Features based on lag and rolling computations are calculated solely from past observations in order to preserve causality.

Model hyperparameters, including the polynomial degree k, the regularization strength α, and the number of selected features k_{best}, are tuned using time-aware cross-validation. Model selection is based on the mean absolute error evaluated on the validation folds. After identifying the optimal configuration for each city, the final model is retrained on the full training dataset.

Inference is performed via recursive multi-step forecasting on the test period by iteratively updating lagged features with previous predictions, thereby simulating real-world deployment conditions. Predictions are aggregated by city, rounded to integer values, and exported as a submission-ready CSV file.

Despite the transparency and easy interpretation of the results of the polynomial regression baseline, its forecast capability is not robust for dengue predictions. Despite the incorporation of seasonal encoding, lag features, and strong regularization, the models struggled to effectively capture the non-stationary dynamics associated with dengue. For the two cities, the optimal solution corresponds to a low Degree Polynomial, indicating higher degree series cause instability without any appreciable performance advantages. This provides the motivation for the use of more expressive models. Examples include gradient boosted decision trees and recurrent neural networks, which are more suited to modeling interactions through time and non-linear feature dependencies.

Table 1: Best Hyperparameters and Cross-Validation MAE Results

City	Degree	α	k_{best}	CV MAE
San Juan	1	10	6	9.1901
Iquitos	1	100	6	4.7527

Table 2: Leaderboard Performance on Test Set

Metric	Value
Submission MAE	27.1875
Leaderboard Rank	4798

3.2 Referencing Prior Works: Ensemble of Models (XGBoost, LSTM, and SVM)

Our ensemble strategy builds upon the multi-modal framework proposed by Sebastianelli et al. (2024)[2], which leverages diverse machine learning architectures to capture the complex spatiotemporal dynamics of dengue transmission. Following the study's emphasis on integrating heterogeneous data clusters—including climatic, vegetation, and socio-economic variables—we developed a robust feature engineering pipeline. This pipeline incorporates cyclical time encodings (sine/cosine transforms of the week index), interaction features between temperature and humidity, and Exponential Weighted Moving Averages (EWMA) to smooth noise in environmental signals.

To ensure model parsimony and focus on categorical importance, we employed a category-balanced feature selection mechanism using XGBoost-based importance scores. Furthermore, we addressed the challenge of abrupt incidence spikes (outliers) by implementing peak-period oversampling and applying exponential decay sample weights, giving higher priority to more recent epidemiological trends.

The core of our predictive engine is a heterogeneous ensemble comprising XGBoost, LSTM, and SVM. This architecture, inspired by the literature, uses LSTM to handle sequential dependencies, while XGBoost and SVM capture high-dimensional non-linear patterns. The final prediction is a weighted aggregation of these base learners:

$$\hat{y}_{t+1} = \Phi\left(h_{\text{XGB}}(X_t), h_{\text{LSTM}}(X_t), h_{\text{SVM}}(X_t)\right) \tag{1}$$

Where Φ represents the aggregation function (meta-learner), and $h_i(X_t)$ denotes the individual hypothesis of each base model at time t.

While the reference study focused on one-month-ahead estimates at the state level, we adapted this methodology to a weekly city-level horizon. By utilizing advanced time-series cross-validation (rolling-window) and carefully tuned hyperparameters, our ensemble successfully mitigated data volatility. This approach yielded a final competition MAE score of **23.3149**.

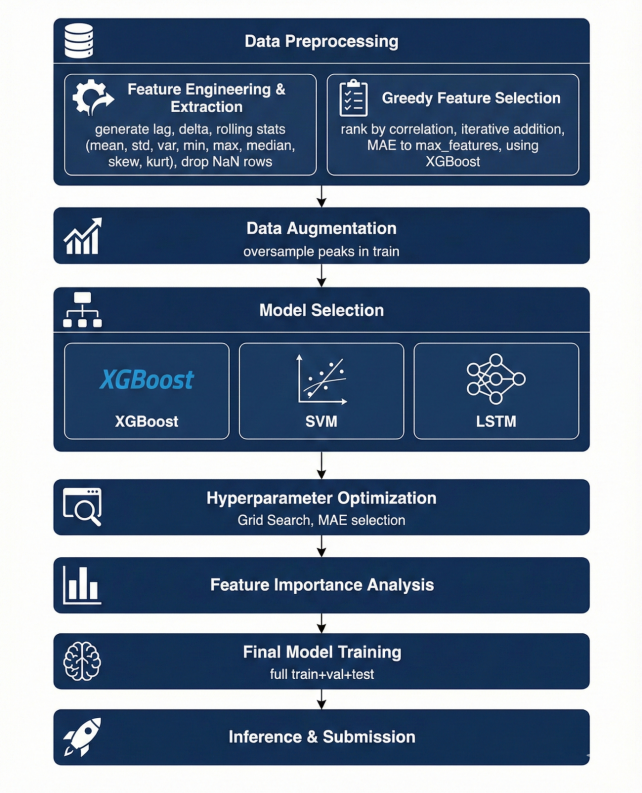

Figure 1: Architecture ensemble of models and its feature engineering workflow.

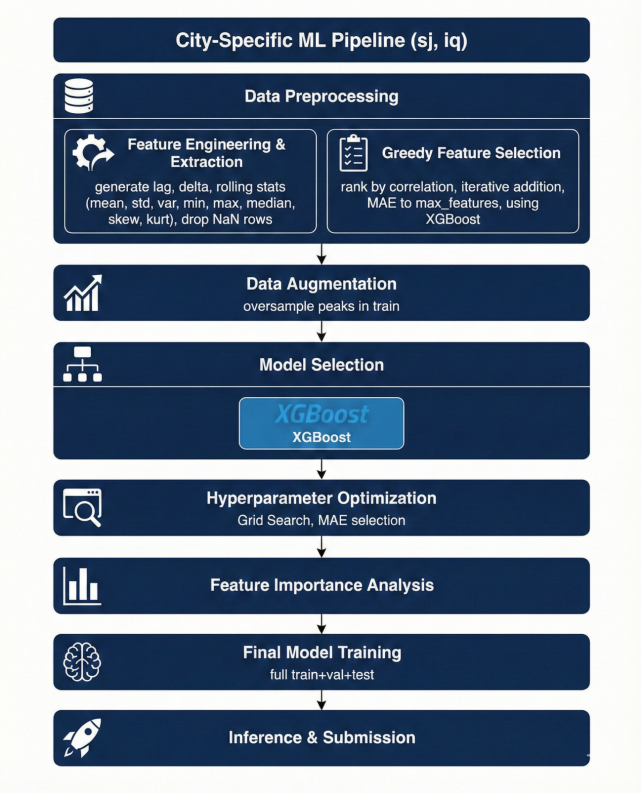

Figure 2: Proposed XGBoost workflow including feature engineering, greedy selection, and training stages.

3.3 XGBoost Regression

Our revised approach shifts the focus from model ensemble diversity to exhaustive feature-space optimization using an XGBoost-centric pipeline, as illustrated in Figure 2. We significantly expanded the feature set by introducing higher-order rolling statistics, including skewness and kurtosis, alongside deltas and multi-period lags for all climate variables. To manage the resulting high-dimensional space, we implemented a greedy forward feature selection strategy. Unlike the importance-based pruning used in our previous ensemble, this iterative method ranks features by their initial correlation and retains only those that provide a measurable reduction in validation Mean Absolute Error (MAE), effectively capping the selection at the top 300 predictors to ensure model parsimony and prevent overfitting.

To address the inherent class imbalance during epidemic peaks, we employed a targeted oversampling strategy for high-incidence rows, forcing the gradient boosting trees to better capture the specific climatic triggers of dengue outbreaks. The dataset was partitioned into a 65/15/20 split for training, validation, and testing, followed by a localized grid search to optimize city-specific hyperparameters. For inference, a "training tail" strategy was utilized to maintain temporal continuity for lag and rolling calculations. The alignment between the model's output and the ground truth is depicted in Figure 3, where predicted values are plotted against actual cases. This refined pipeline achieved a leaderboard MAE of **22.6202**, demonstrating that rigorous feature selection and imbalance handling can outperform more complex heterogeneous ensembles.

3.4 Predicting Dengue Cases Ahead of Time

This approach diverges from the prior XGBoost implementation by incorporating multi-horizon forecasting, evaluating predictions from $t + 1$ to $t + 12$ weeks ahead to select the optimal horizon per city, as detailed in Table 3. Instead of single-step forecasting, this method loads pre-selected features from previous XGBoost model and employs horizon-specific target shifting.

While this strategy prioritizes extended temporal forecasting for public health planning, it yielded a higher submission Mean Absolute Error (MAE) of 28.4351 compared to the previous 22.6202, highlighting the complexities of long-term prediction. Despite the drop in overall leaderboard performance, the model retains robust preprocessing, including lags, rolling statistics, deltas, and oversampling.

The evaluation on validation and test sets reveals distinct behaviors per city. In **San Juan**, the selection of $t + 12$ as the optimal horizon is driven by the minimum validation MAE (10.1200). However, this horizon exhibits the highest error on the test set (25.8883), suggesting potential overfitting to the validation split or a shift in data distribution at that specific lag.

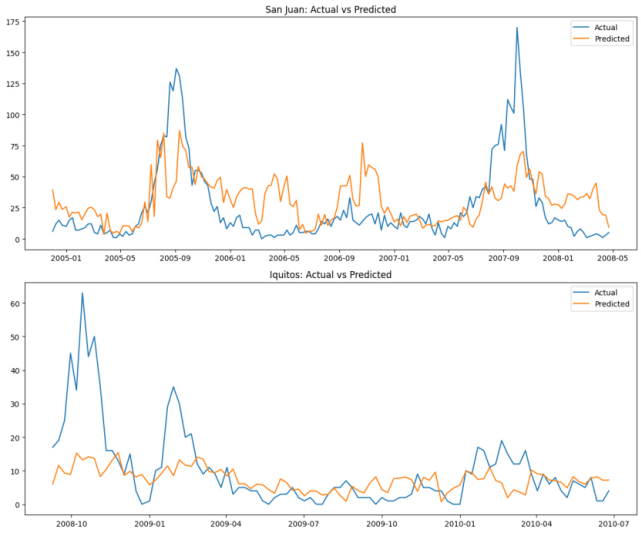

Figure 3: Visualization of actual vs. predicted dengue cases on the test sets for San Juan and Iquitos.

Table 3: Horizon Comparison Summary for San Juan and Iquitos

City	Horizon ($t+n$)	Validation MAE	Test MAE
San Juan	1	12.7187	20.7272
	2	14.7684	21.6136
	4	12.9712	20.9721
	8	10.6847	20.3510
	12	**10.1200**	25.8883
Selected Best Horizon: t+12			
Iquitos	1	8.8262	7.4369
	2	**7.3934**	7.9215
	4	9.2280	7.9130
	8	7.9530	8.2729
	12	7.7922	10.1311
Selected Best Horizon: t+2			

Conversely, **Iquitos** shows higher stability; the chosen $t + 2$ horizon (7.3934 Val MAE) generalizes well to the test set (7.9215 Test MAE). The discrepancy between validation and test performance in San Juan is a primary contributor to the increased submission MAE, as the model optimized for $t + 12$ fails to maintain its accuracy on unseen data compared to the $t + 8$ or $t + 1$ horizons.

3.5 Other Models & Comparison

In this section, we provide an extensive comparison of each of the implemented approaches. Table 4 summarizes the performance of these methods, including city-specific validation errors, final test scores on the DrivenData evaluation environment, and the corresponding leaderboard rankings.

Discussion of Results: Table 4 indicates notable performance differences among the evaluated models. The tuned **XGBoost** model with lag and rolling statistical features achieved the best overall performance, obtaining the lowest leaderboard MAE of 22.62 and a strong rank of 574.

This result confirms the effectiveness of tree-based ensemble methods in modeling non-linear relationships between climatic variables and dengue case counts. The **ensemble model (XGBoost + LSTM + SVM)** delivered a competitive performance and outperformed the polynomial baseline as well as the individual LSTM and SVM models; however, it did not surpass the standalone XGBoost model. This suggests that, while heterogeneous ensemble techniques can be effective in reducing variance, the inclusion of relatively weaker learners may limit the overall gain when aggregation weights are not optimally tuned.

In particular, the **PCA-based XGBoost** variant exhibited the weakest performance among learning-based methods. Based on both validation and leaderboard results, it can be inferred that linear dimensionality reduction eliminated important city-specific and non-linear information required for accurate dengue forecasting.

The **LSTM** and **SVM** models achieved moderate performance, demonstrating their ability to learn temporal and non-linear patterns, although they consistently underperformed compared to XGBoost-based approaches. Overall, all machine learning models, except for the PCA-based method, outperformed the k-polynomial regression baseline, emphasizing the importance of expressive and flexible models for capturing the complex and non-stationary dynamics of dengue transmission.

Table 4: Performance Summary (Validation vs. Leaderboard). Sorted by Test Score.

Model	SJ (Val)	IQ (Val)	Test (MAE)	Rank
XGBoost	**9.16**	11.37	**22.62**	**574**
Ensemble (XGB+LSTM+SVM)	17.89	**5.40**	23.31	1463
SVR	11.57	11.08	26.54	3785
k-Polynomial Baseline	9.19	4.75	27.19	4798
Ahead-of-time XGB	18.95	6.79	28.44	5158
LSTM	17.92	5.49	28.60	5797
PCA + XGBoost	18.13	8.40	32.28	7312

4 Feature Importance Analysis

The evaluation of feature importance for the San Juan (SJ) and Iquitos (IQ) models reveals significant differences in the environmental drivers of dengue fever. In San Juan, the model is heavily dominated by the diurnal temperature range, contributing nearly 49% of the predictive weight. This suggests that long-term deviations in daily temperature fluctuations are the primary indicators for outbreaks in this coastal urban environment. In contrast, the Iquitos model is more ecologically sensitive, with vegetation indices (NDVI) and precipitation patterns showing higher relevance, reflecting the city's

location within the Amazon rainforest where humidity and jungle density play a larger role.

The following tables (Table 5 and Table 6) summarize the top five features for each city.

Table 5: Top 5 Feature Importances for San Juan (SJ).

Feature Name	Category	Importance
station_diur_temp_rng_c_rolling_skew_52	Temp. Range	**0.4915**
reanalysis_air_temp_k_rolling_var_20	Temperature	0.0993
precipitation_amt_mm_rolling_std_52	Precipitation	0.0840
station_diur_temp_rng_c_rolling_std_20	Temp. Range	0.0752
station_min_temp_c_lag_36	Temperature	0.0732

Table 6: Top 5 Feature Importances for Iquitos (IQ).

Feature Name	Category	Importance
ndvi_sw_rolling_kurt_52	Vegetation (NDVI)	**0.3318**
reanalysis_max_air_temp_k_rolling_max_28	Temperature	0.1154
reanalysis_precip_amt_kg_per_m2_rolling_kurt_52	Precipitation	0.0953
ndvi_nw_rolling_skew_28	Vegetation (NDVI)	0.0659
station_min_temp_c_rolling_min_4	Temperature	0.0572

5 Conclusion

In this term project we investigated weekly dengue case forecasting using the DrivenData DengAI benchmark in accordance with the course requirements. A high-degree polynomial regression model was implemented as a baseline, revealing limited capability in capturing the non-linear and non-stationary dynamics of dengue time-series data despite extensive regularization and feature engineering.

Several machine learning approaches were then evaluated, including SVR, LSTM, XGBoost, and a heterogeneous ensemble. Among all methods, the tuned XGBoost model achieved the best performance with a leaderboard MAE of 22.62 and a rank of 574, outperforming both the polynomial baseline and the ensemble model. This result highlights the effectiveness of tree-based ensemble methods and shows that increased model complexity does not necessarily yield performance gains when weaker learners are included.

As required by the project, a second model was implemented for ahead-of-time forecasting using only current information. The optimal prediction horizons were identified as t+12 weeks for San Juan and t+2 weeks for Iquitos, yielding a leaderboard MAE of 28.44 and illustrating the trade-off between early prediction and accuracy.

Overall, the results demonstrate the importance of expressive models, robust feature engineering, and city-specific modeling strategies for reliable dengue forecasting.

References

[1] [n. d.]. DengAI: Predicting Disease Spread. DrivenData competition. Accessed: 2026-01-12.

[2] A. Sebastianelli, D. Spiller, R. Carmo, et al. 2024. A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports* (2024). doi:10.1038/s41598-024-52796-9