# DengAI:

# Predicting Disease Spread

Bharat Jain, Soorya K S, Usman Akhinyemi

# 01

# Problem Statement

# Problem we are trying to solve!

According to the World Health Organization, dengue fever is one of the top ten global health threats – it's also the most rapidly spreading.

*"Each year, up to **400 million** people are infected by a dengue virus. Approximately **100 million people get sick** from infection, and **40,000 die from severe dengue**."*

*Centers for Disease Control and Prediction*

So, there's an urgent demand for effective strategies to predict numbers of dengue cases and mitigate their impact on global health.

# Potential applications of the solution?

Developing an accurate model to predict the number of dengue cases to

- Enables better healthcare planning and resource allocation
- Target mosquito control efforts
- Establish early warning systems
- Guide for future research and policy decisions
- Strengthen global health security efforts

# Impacts of the solution?

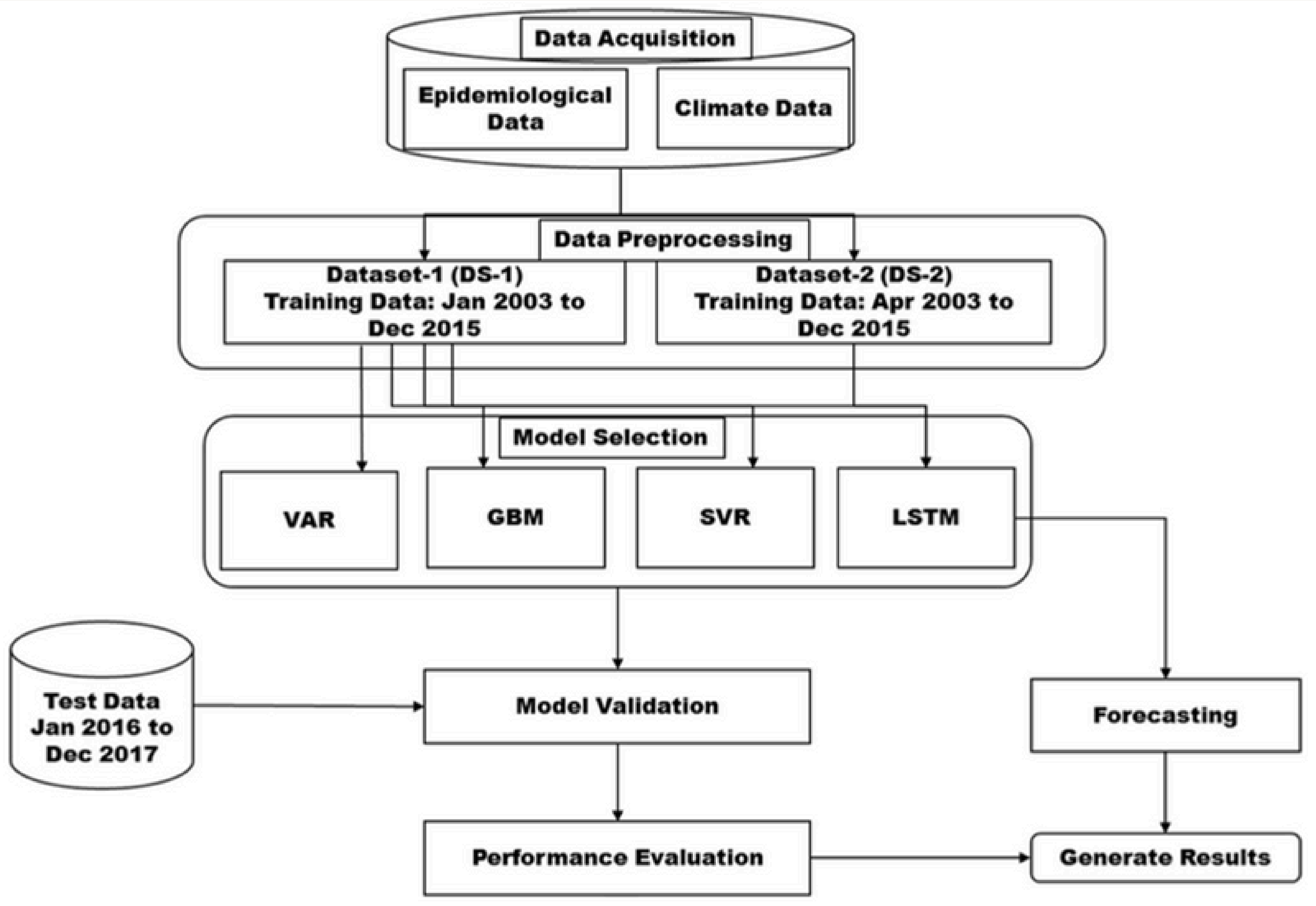
- Mitigation of global public health threats, including dengue fever
- Improved public health outcomes through accurate forecasts, leading to proactive measures and resource allocation.
- Cost savings through the implementation of smarter, more efficient strategies for disease control.
- Advancements in disease preparedness, future vector-borne diseases

# 02

# Literature Survey

# Paper:1

"S. G. Kakarla et al., "Weather integrated multiple machine learning models for prediction of dengue prevalence in India," Int. J. Biometeorol., vol. 67, no. 2, pp. 285–297, 2023."




**Key Points:**

- Kerala's coastal location increases dengue risk due to factors like rain and humidity.
- Lag features (using past data) were used to predict future dengue cases accurately.
- Different models were combined for forecasting.
- Urban areas in Kerala had higher dengue cases, highlighting the need for targeted interventions.

# Paper:1

"S. G. Kakarla et al., "Weather integrated multiple machine learning models for prediction of dengue prevalence in India," Int. J. Biometeorol., vol. 67, no. 2, pp. 285–297, 2023."

**Features Used**

Weather Variables
Lagged Variables
Dengue Cases

| Models Used | RMSE | Coeffcient of determination (r^2) | Variance Explained |
|---|---|---|---|
| 1) Vector Auto Regression (VAR) model | 0.572 | 0.67 | - |
| 2) Support Vector Regression (SVR) | 0.447 | 0.8 | 90% |
| 3) Generalized Boosted Regression Model (GBM) | 1.65 | 0.36 | 36% |
| 4) Long Short-Term Memory (LSTM) model | 0.345 | 0.86 | 86% |

# Paper:2

“Predicting Dengue Fever Outbreaks,” Gregcondit.com. [Online]. Available: https://www.gregcondit.com/projects/dengue-fever. [Accessed: 11-May-2024].

1) Corrected date anomalies and examined weather patterns correlation with dengue cases

2) Selected key weather variables based on domain knowledge and exploratory analysis.

3) LSTM neural networks for predicting Dengue outbreaks but found limitations in model performance due to dataset size and complexities.

4) Utilized lagged features with Random Forest Regressor to incorporate time dependencies in predictions.

5) Walk Forward Validation: Implemented a validation strategy that progresses through time to validate models effectively without violating the time order of data

# Paper:2

“Predicting Dengue Fever Outbreaks,” Gregcondit.com. [Online]. Available: https://www.gregcondit.com/projects/dengue-fever. [Accessed: 11-May-2024].

Models Used:
LSTM
RandomForestTree

Performance Metrics
- Mean Absolute Error (MAE) : 24

Best predictions are using Random Forest Regressors with 3 weeks of lagged features


# 03

# Dataset

# Collection

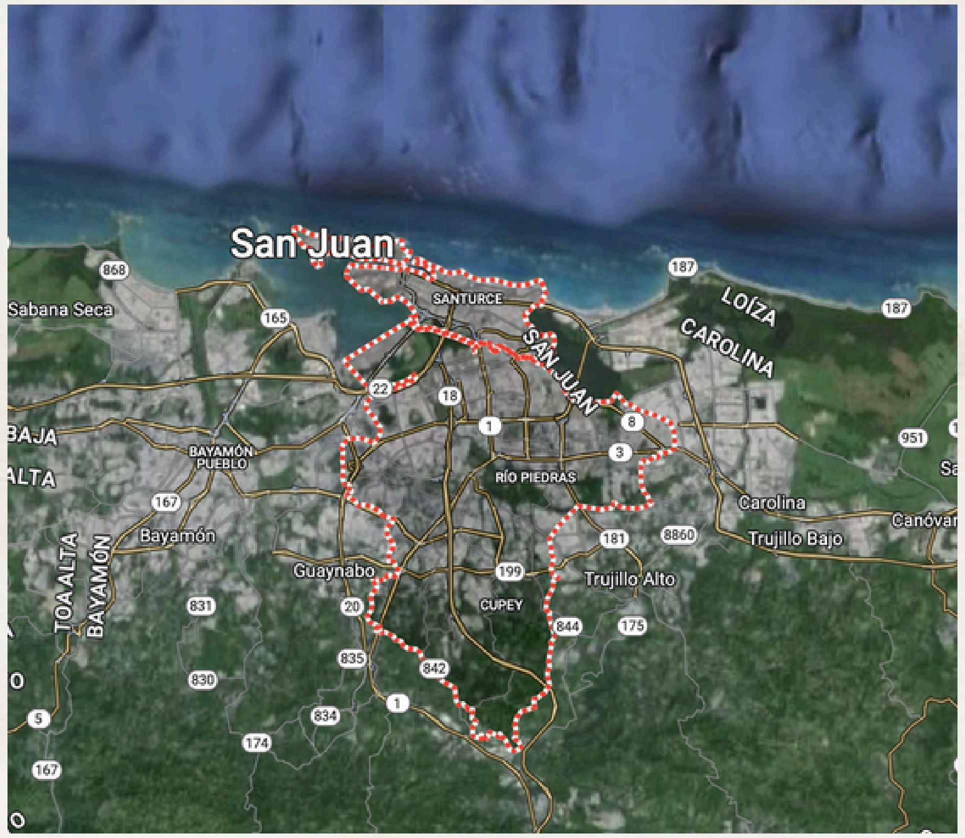
Dengue surveillance data is provided by

- The U.S. **Centers for Disease Control and prevention**
- Department of Defense's Naval Medical Research Unit and the Armed Forces Health Surveillance Center, in collaboration with the Peruvian government and U.S. universities.

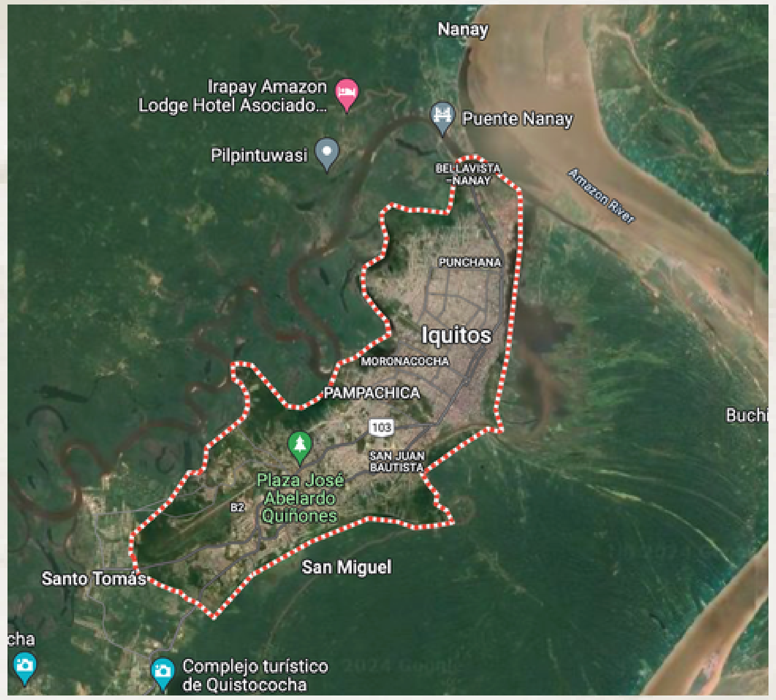
Environmental and climate data is provided by

- The **National Oceanic and Atmospheric Administration (NOAA)**, an agency of the U.S. Department of Commerce.

# San Juan and Iquitos




- San Juan capital of Puerto Rico, located at the northern coast of the island, on the Atlantic Ocean.





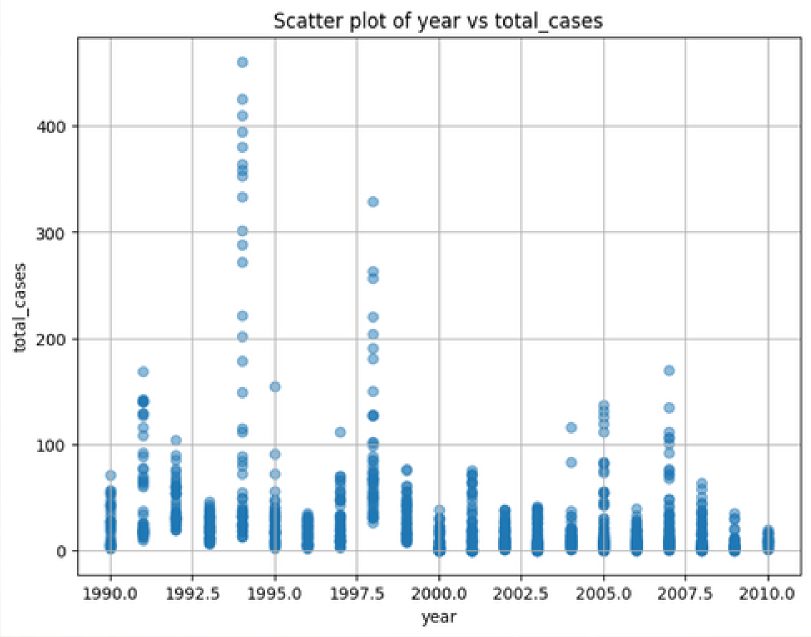
- Iquitos, capital of Peru's Maynas Province and Loreto Region is the largest metropolis in the Peruvian Amazon, as well as the ninth-most populous city in Peru.

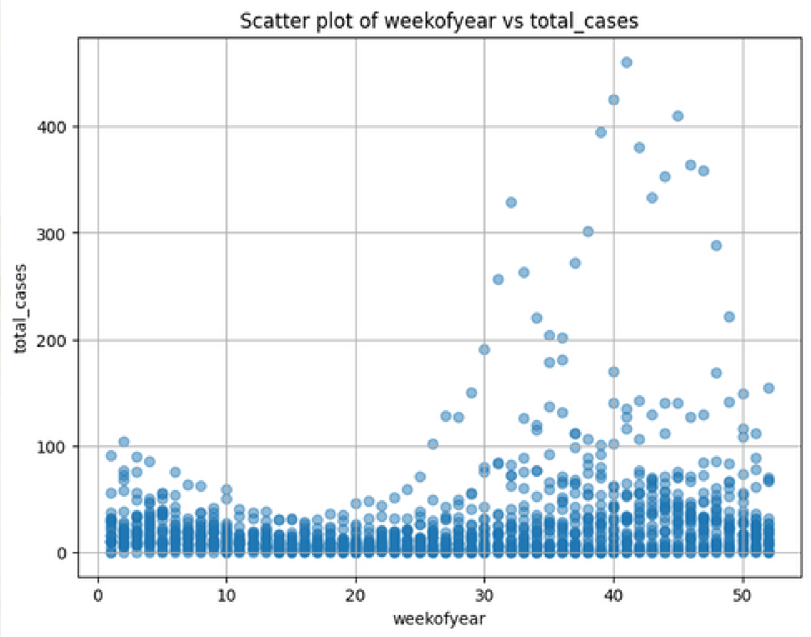
# Nature of the Dataset

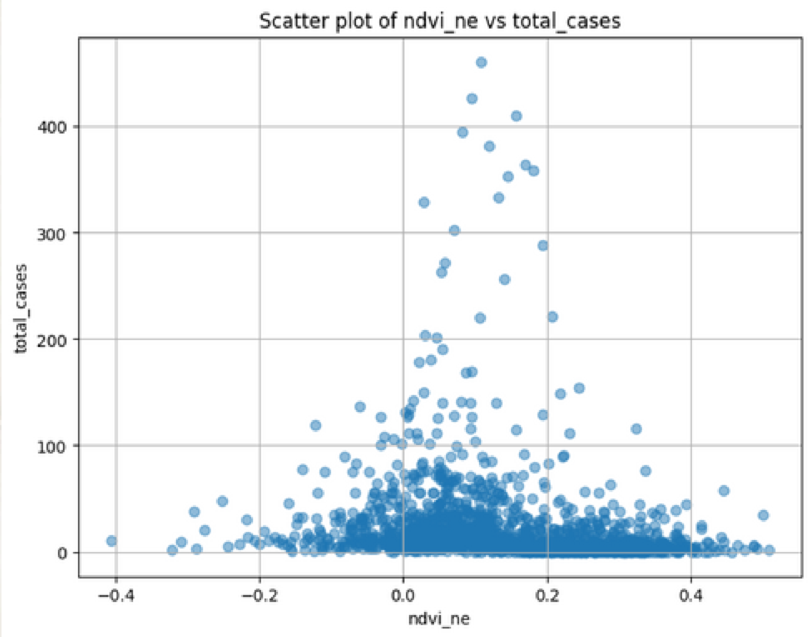
| Classes | Features |
| --- | --- |
| time_group | year, weekofyear, week_start_date, weekofyear_fixed |
| vegetation_index_group | ndvi_ne, ndvi_nw, ndvi_se, ndvi_sw |
| precipitation_group | precipitation_amt_mm, reanalysis_precip_amt_kg_per_m2, reanalysis_sat_precip_amt_mm, station_precip_mm |
| temperature_group | reanalysis_air_temp_k', 'reanalysis_avg_temp_k', 'reanalysis_max_air_temp_k', 'reanalysis_min_air_temp_k', 'station_avg_temp_c', 'station_max_temp_c', 'station_min_temp_c |
| humidity_group | reanalysis_dew_point_temp_k, reanalysis_relative_humidity_percent, reanalysis_specific_humidity_g_per_kg |

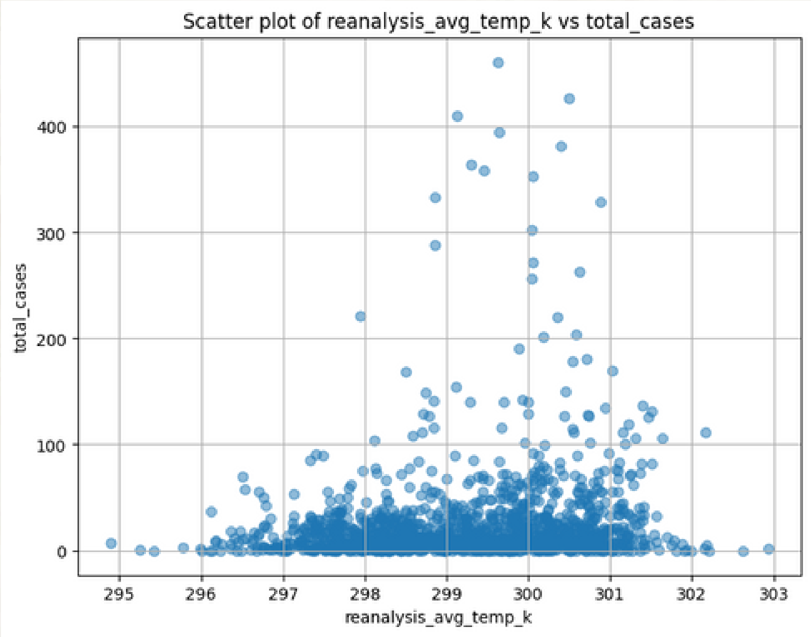
**Total number of features: 23**
**Total train data points: 1456**
**Total test data points: 416**

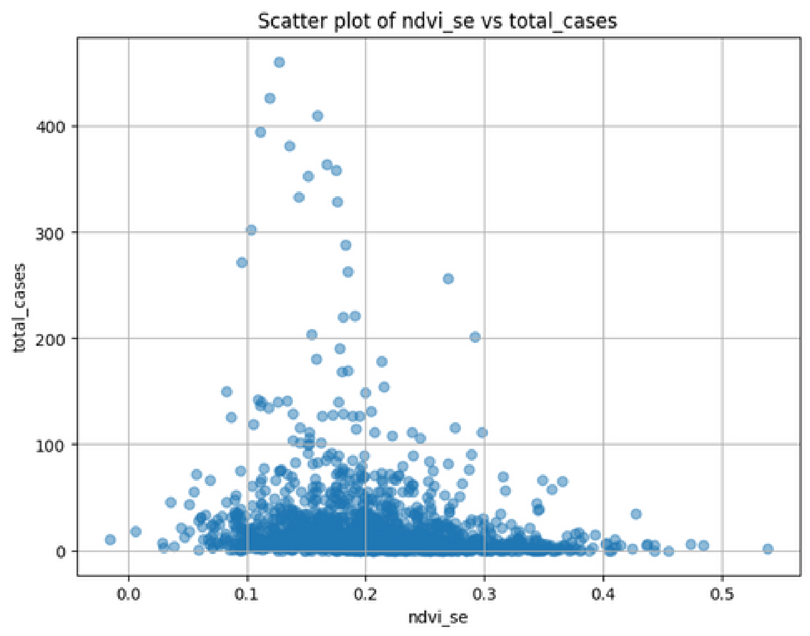
# Visualization

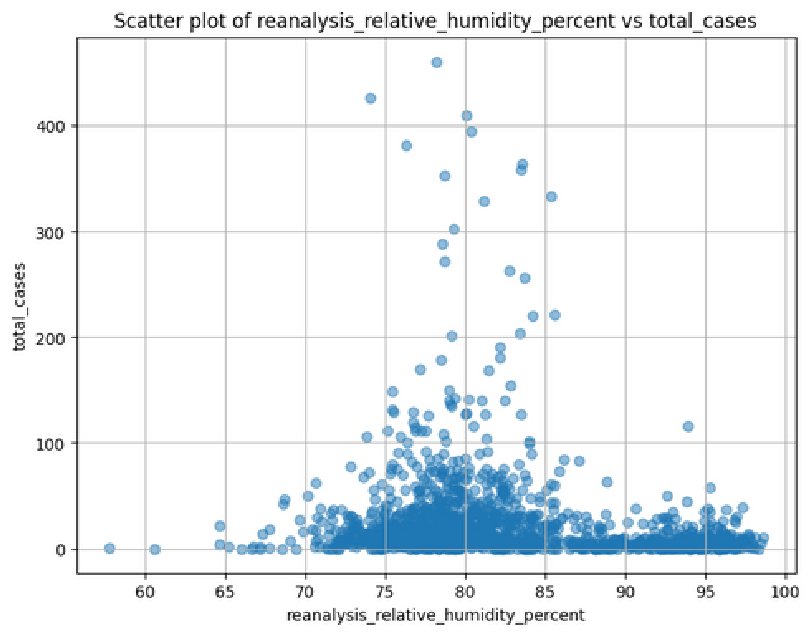


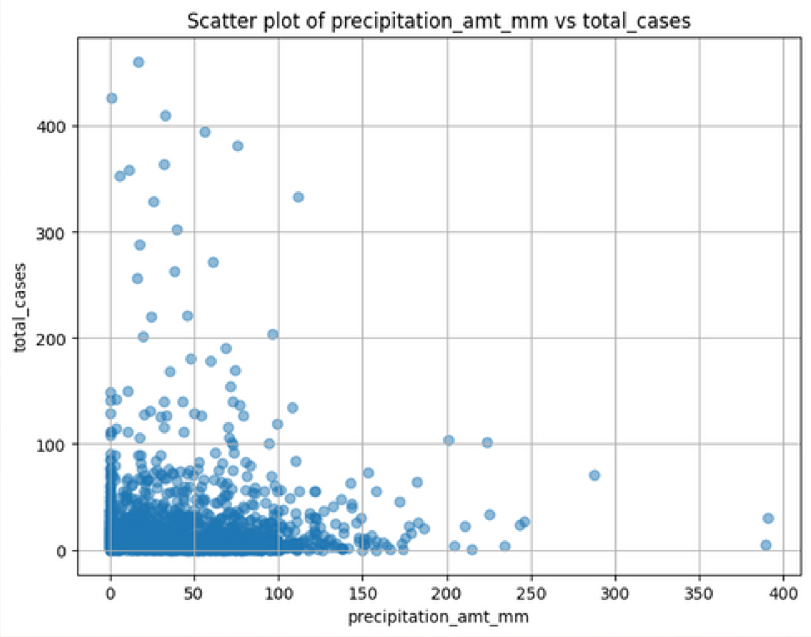



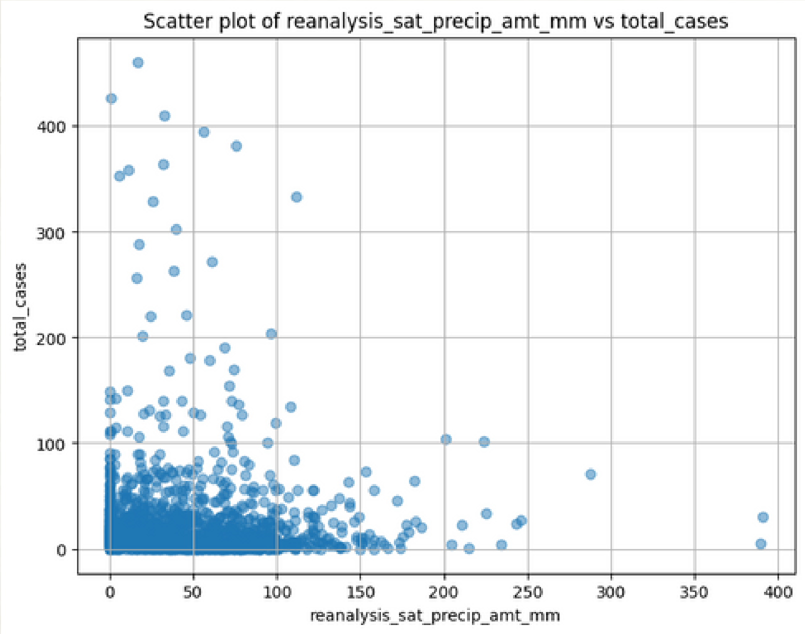



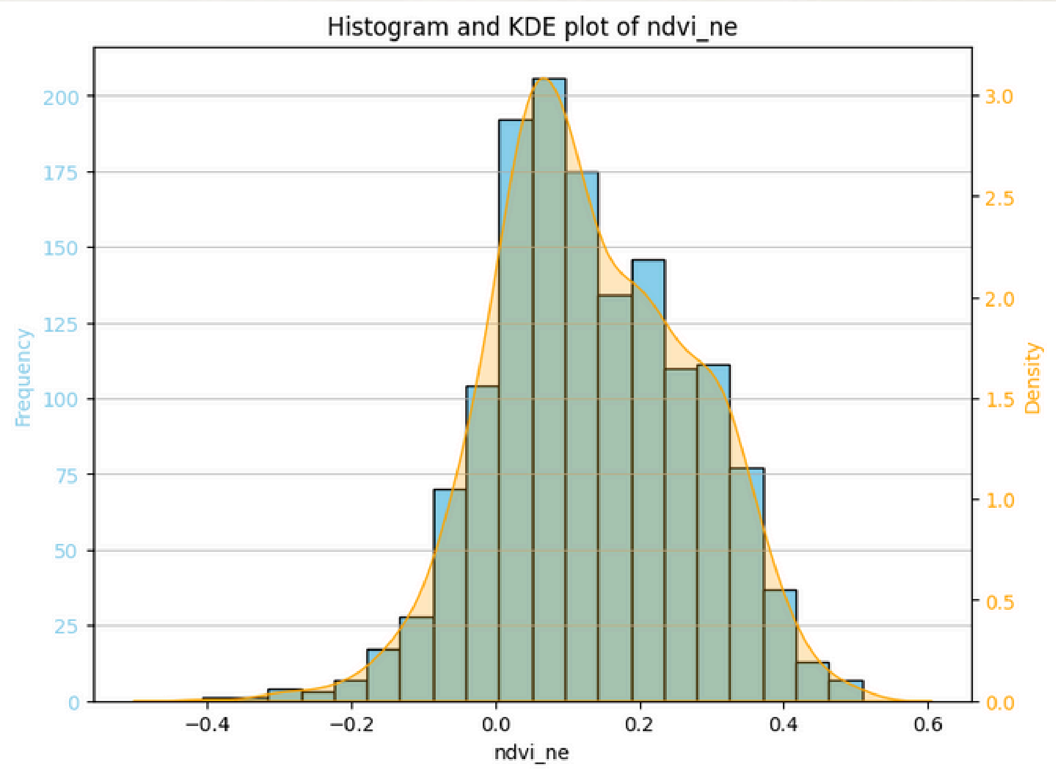



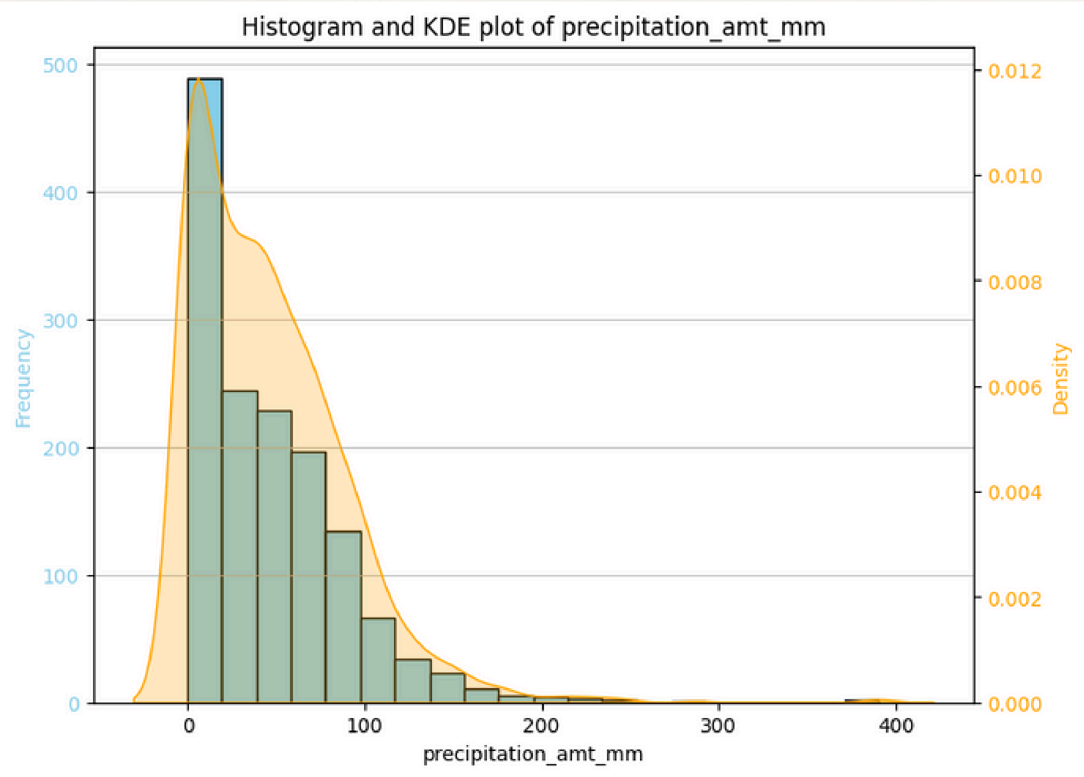



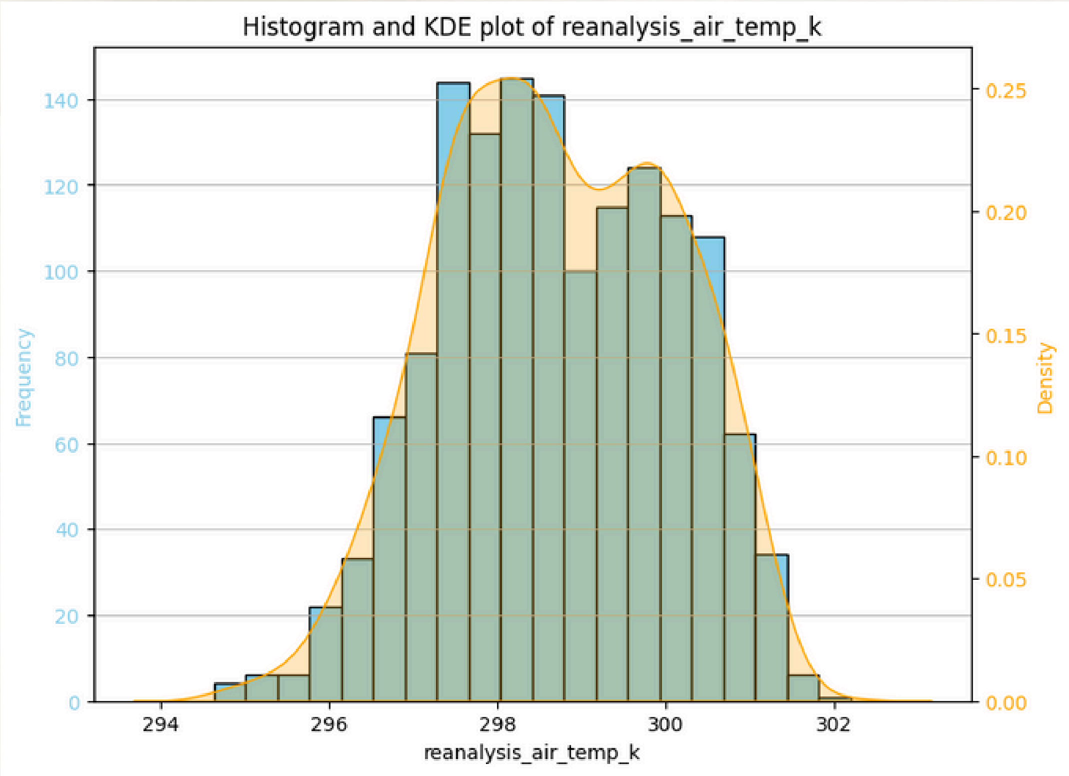



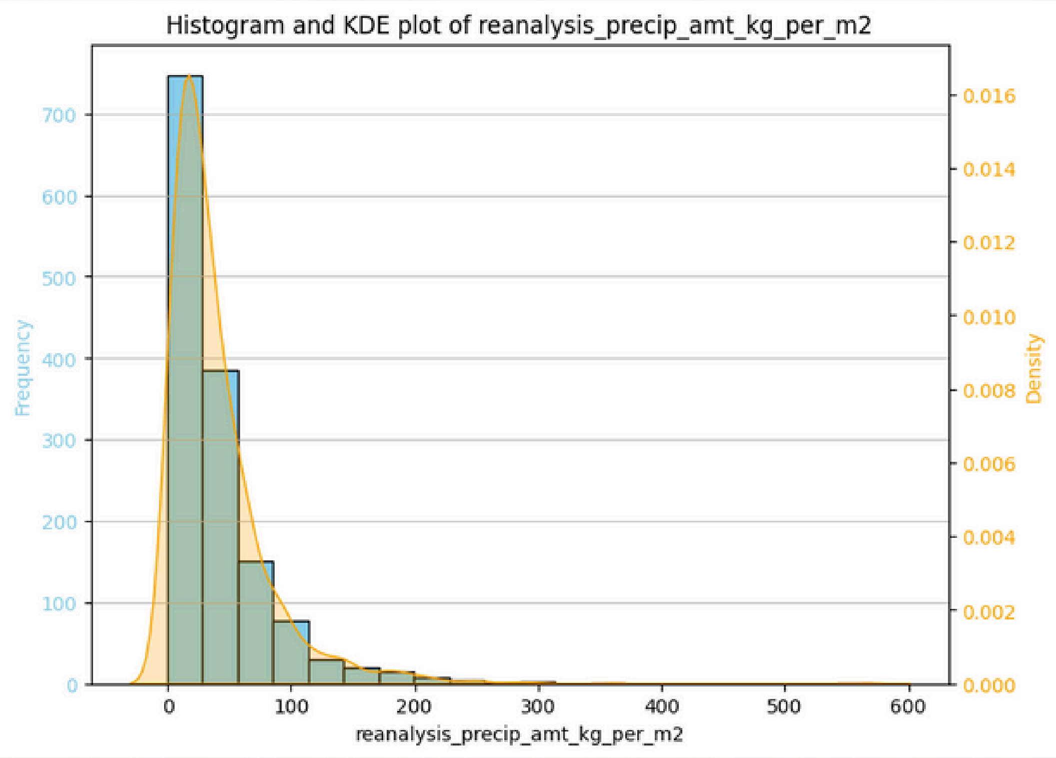



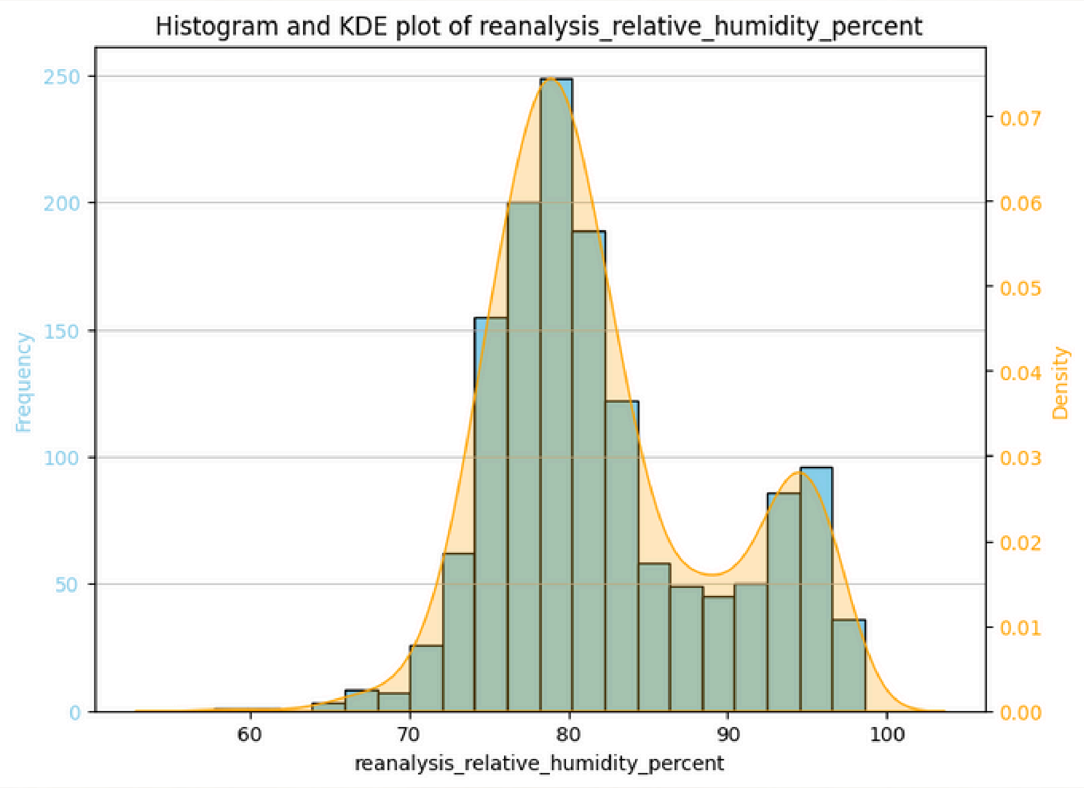
# Skewed Distribution of Dataset

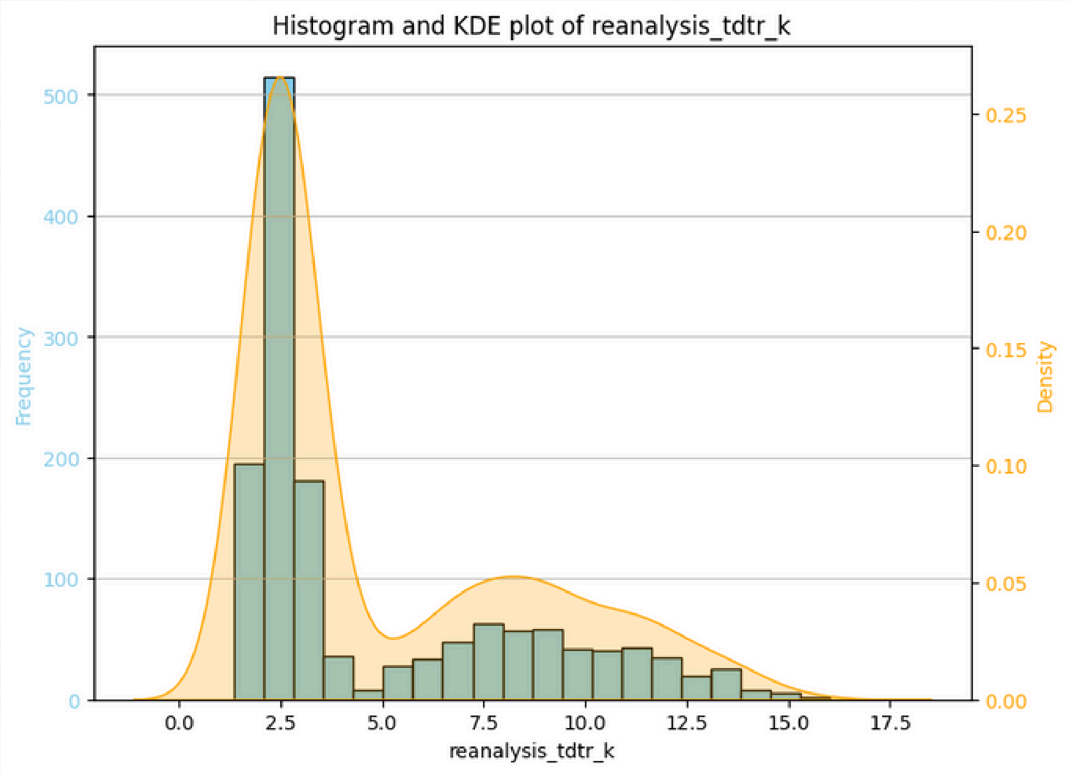


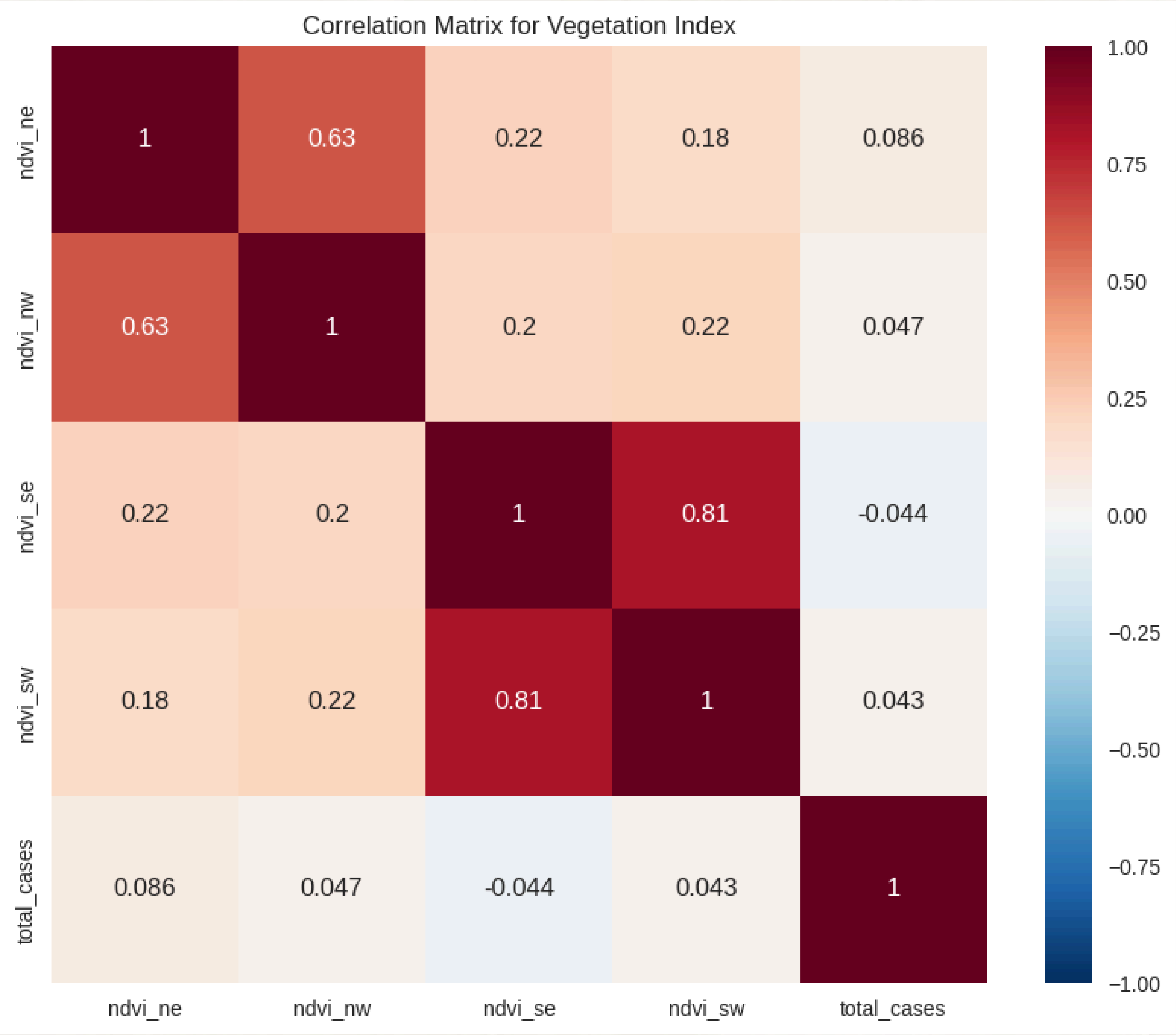



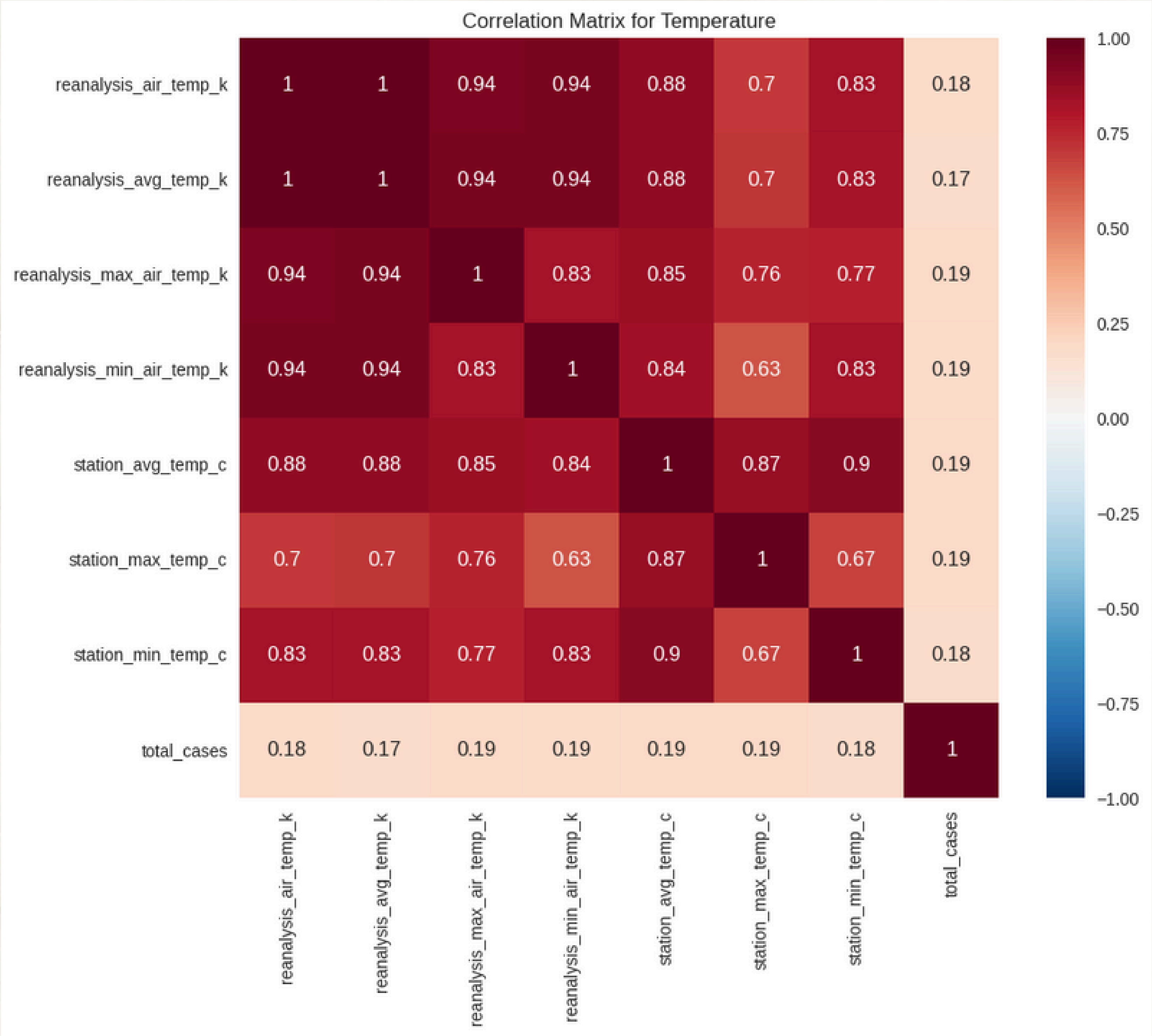



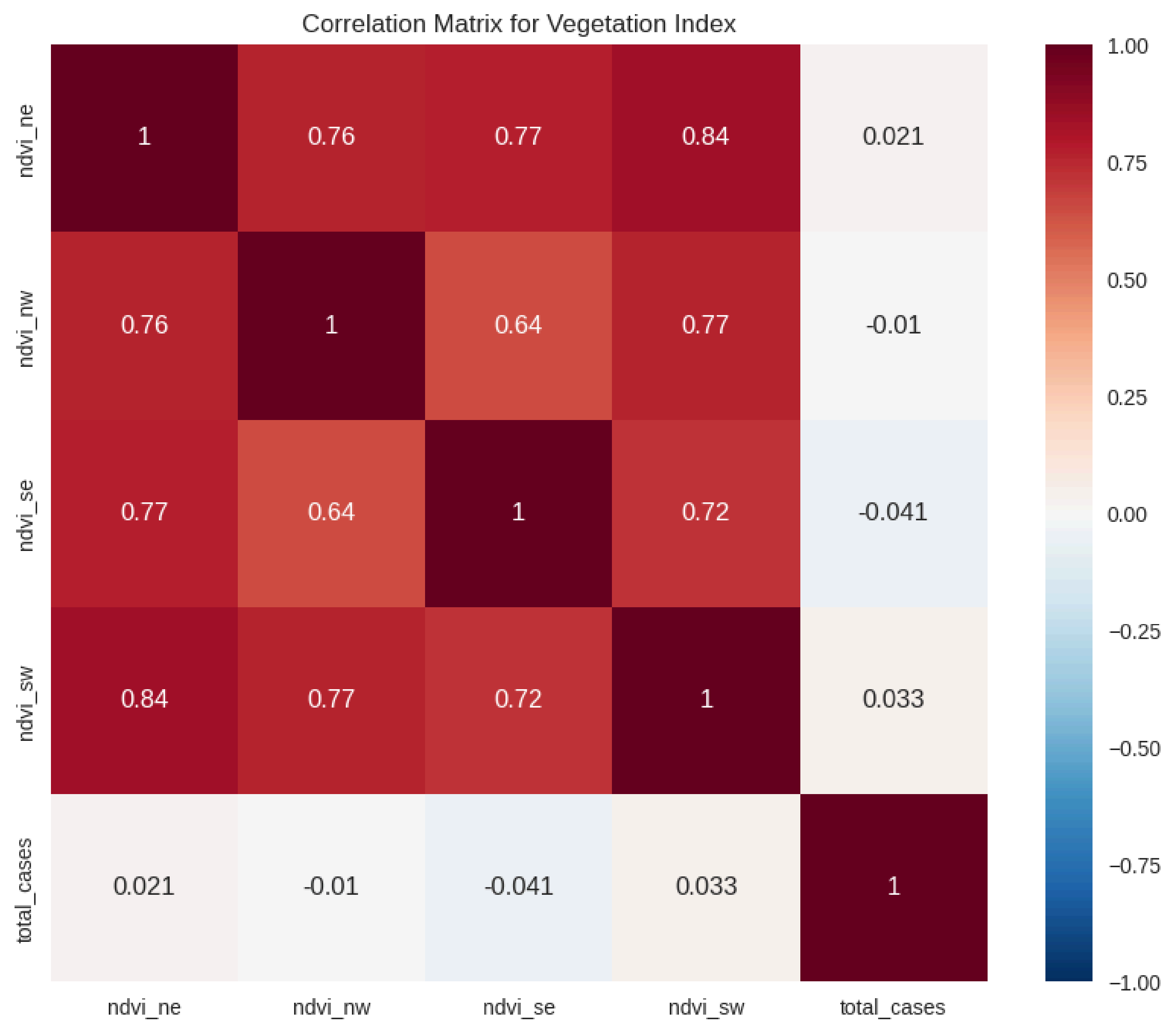
# Correlation amongst groups and with Target

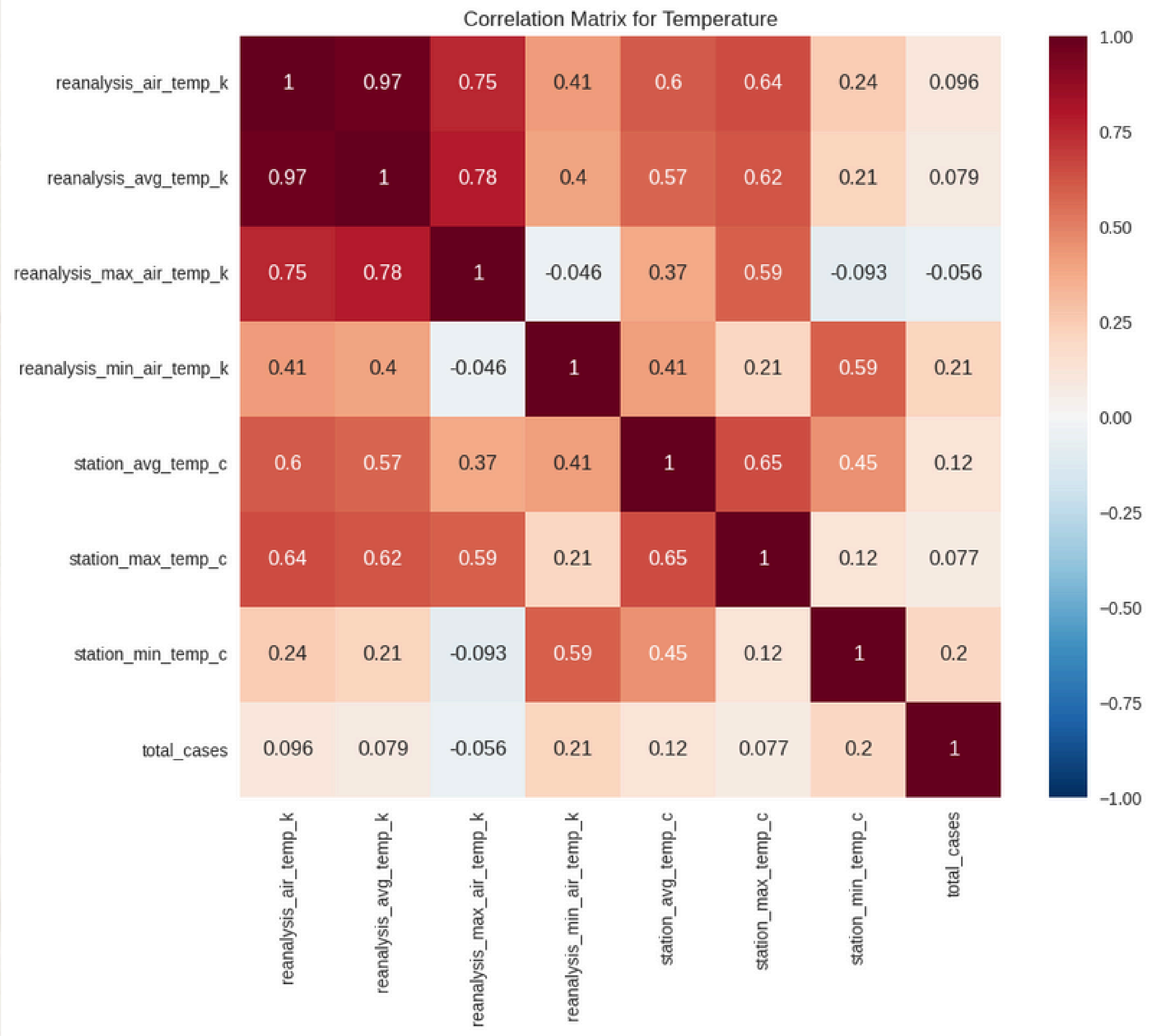



San Juans

# Correlation amongst groups and with Target




Iquitos

# Results on PCA on the Test

| Models | MAE (without PCA) | MAE (with PCA) |
|---|---|---|
| Elastic_Net | 26.3606 | 27.6322 |
| Bayesian_Regression | 26.5649 | 27.5889 |
| DummyRegressor | 27.6827 | 27.6827 |
| ExtraTressRegressor | 26.0817 | 30.8077 |
| HuberRegressor | 27.7428 | 28.7933 |
| LassoRegressor | 26.7428 | 27.6298 |

| Models | MAE (without PCA) | MAE (with PCA) |
|---|---|---|
| LassoLars | 26.7428 | 27.6298 |
| OrthogonalMatchingPursuit | 27.0986 | 27.6514 |
| Ridge | 26.6875 | 27.6418 |
| RandomForest Regressor | 26.6442 | 30.1755 |

# 04

# Feature Preprocessing

## For each city:

- **Fixing Weeks:**
  - Identify years where the 53rd week exists. This is crucial because not all years have a 53rd week.
  - Identify years with the 53nd week as the first week and adjust week numbers by incrementing with 1.

```
Year: 1993
```

| | city | year | weekofyear_fixed | weekofyear | week_start_date |
|---|---|---|---|---|---|
| **139** | sj | 1993 | 1 | 53 | 1993-01-01 |
| **140** | sj | 1993 | 2 | 1 | 1993-01-08 |
| **141** | sj | 1993 | 3 | 2 | 1993-01-15 |

| | city | year | weekofyear_fixed | weekofyear | week_start_date |
|---|---|---|---|---|---|
| **188** | sj | 1993 | 50 | 49 | 1993-12-10 |
| **189** | sj | 1993 | 51 | 50 | 1993-12-17 |
| **190** | sj | 1993 | 52 | 51 | 1993-12-24 |

**For each city,**

- **Handling Missing Values:**
    - Identify columns with missing values, especially for NDVI (Normalized Difference Vegetation Index).
    - Interpolate missing values using linear interpolation method
    - Apply the Climatological Mean of the Day (CMD) method for interpolating climate data

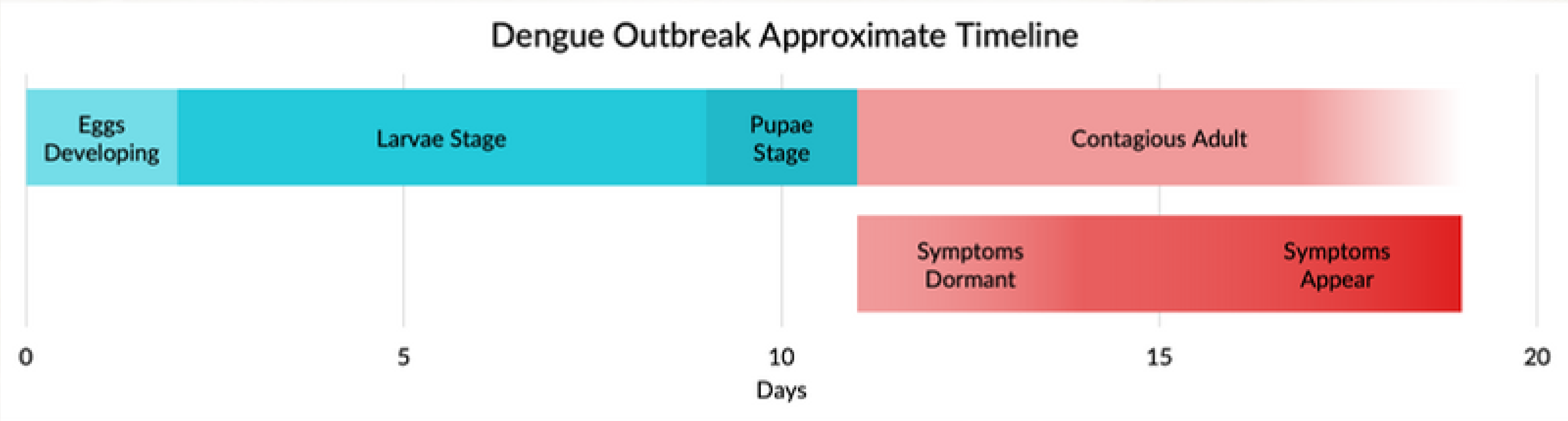
$$V_{est} = \frac{\sum_{j=1}^{T} V_{ij}}{T}$$

    - Vi is the ith day of year j. T is the number of available data for that year. (Narapusetty, et al. Optimal estimation of the climatological mean)

| | |
|---|---|
| station_diur_temp_rng_c | 37 |
| station_avg_temp_c | 37 |
| station_precip_mm | 16 |
| station_max_temp_c | 14 |
| station_min_temp_c | 8 |
| reanalysis_max_air_temp_k | 4 |
| reanalysis_tdtr_k | 4 |
| reanalysis_specific_humidity_g_per_kg | 4 |
| reanalysis_sat_precip_amt_mm | 4 |
| reanalysis_relative_humidity_percent | 4 |
| reanalysis_precip_amt_kg_per_m2 | 4 |
| reanalysis_min_air_temp_k | 4 |
| reanalysis_dew_point_temp_k | 4 |
| reanalysis_air_temp_k | 4 |
| precipitation_amt_mm | 4 |
| reanalysis_avg_temp_k | 4 |
| ndvi_sw | 3 |
| ndvi_se | 3 |
| ndvi_nw | 3 |
| ndvi_ne | 3 |

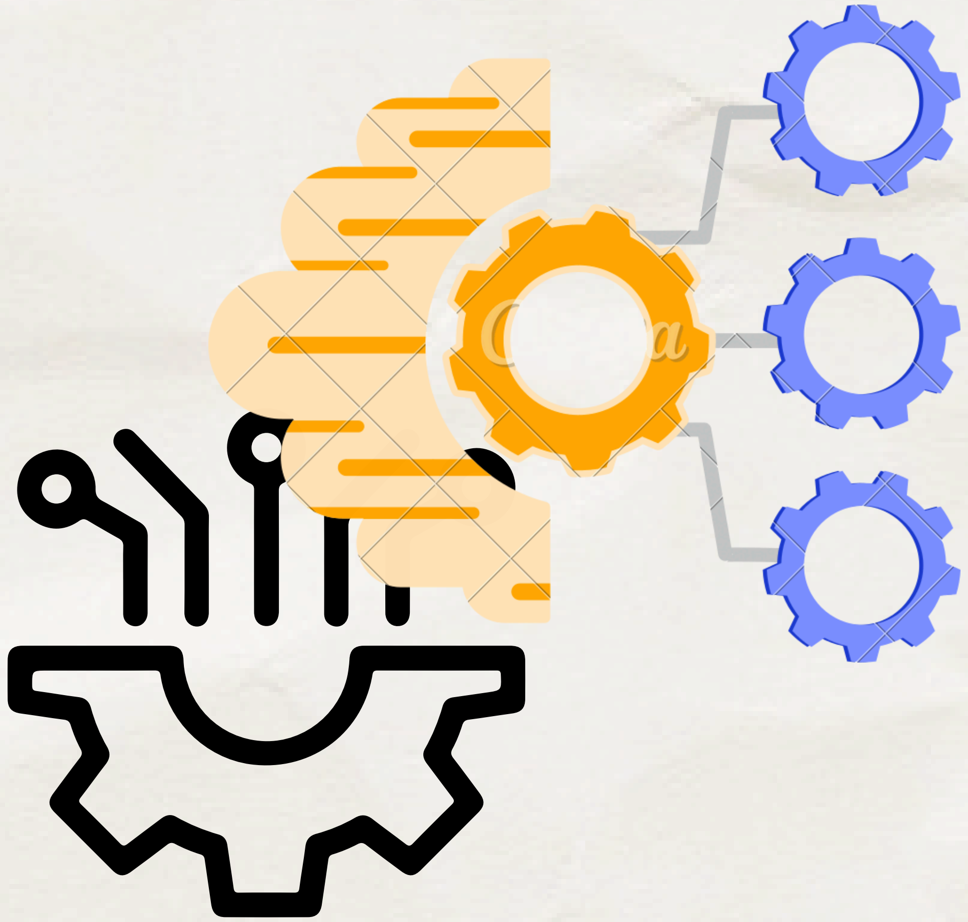
| | |
|---|---|
| week_start_date | 0 |
| ndvi_ne | 0 |
| ndvi_nw | 0 |
| ndvi_se | 0 |
| ndvi_sw | 0 |
| precipitation_amt_mm | 0 |
| reanalysis_air_temp_k | 0 |
| reanalysis_avg_temp_k | 0 |
| reanalysis_dew_point_temp_k | 0 |
| reanalysis_max_air_temp_k | 0 |
| reanalysis_min_air_temp_k | 0 |
| reanalysis_precip_amt_kg_per_m2 | 0 |
| reanalysis_relative_humidity_percent | 0 |
| reanalysis_sat_precip_amt_mm | 0 |
| reanalysis_specific_humidity_g_per_kg | 0 |
| reanalysis_tdtr_k | 0 |
| station_avg_temp_c | 0 |
| station_diur_temp_rng_c | 0 |
| station_max_temp_c | 0 |
| station_min_temp_c | 0 |
| station_precip_mm | 0 |

- Data standardization
- Adding lagg features




Juliano SA, O'Meara GF, Morrill JR, Cutwa MM. Desiccation and thermal tolerance of eggs and the coexistence of competing mosquitoes. Oecologia. 2002;130(3):458–469. doi:10.1007/s004420100811

We've looked at past data and added new features by shifting our climatic information by about 3 weeks



# 05

# Machine Learning Model

# Model Selection: Shortlisting Relevant Models

- PyCaret Python Library was used to do a quick run through
- Ran the function for both cities seprately (similar models were there in top 10)
- Shortlisted the best 10 models for further analysis of performance

| | Model | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|---|
| **et** | Extra Trees Regressor | 25.8376 | 2284.5159 | 46.9393 | 0.1495 |
| **en** | Elastic Net | 27.9897 | 2603.2451 | 49.8333 | 0.0586 |
| **br** | Bayesian Ridge | 28.3362 | 2591.1884 | 49.7721 | 0.0574 |
| **lasso** | Lasso Regression | 28.3876 | 2590.0591 | 49.8091 | 0.0536 |
| **llar** | Lasso Least Angle Regression | 28.3847 | 2591.0750 | 49.8174 | 0.0533 |
| **ridge** | Ridge Regression | 28.9281 | 2606.5684 | 49.9518 | 0.0479 |
| **omp** | Orthogonal Matching Pursuit | 28.1983 | 2589.3407 | 49.8845 | 0.0456 |
| **lr** | Linear Regression | 29.0921 | 2630.4215 | 50.1649 | 0.0407 |
| **huber** | Huber Regressor | 24.3296 | 2792.2081 | 51.4973 | -0.0005 |
| **dummy** | Dummy Regressor | 29.2975 | 2809.7766 | 51.7658 | -0.0142 |

**For San Juan**

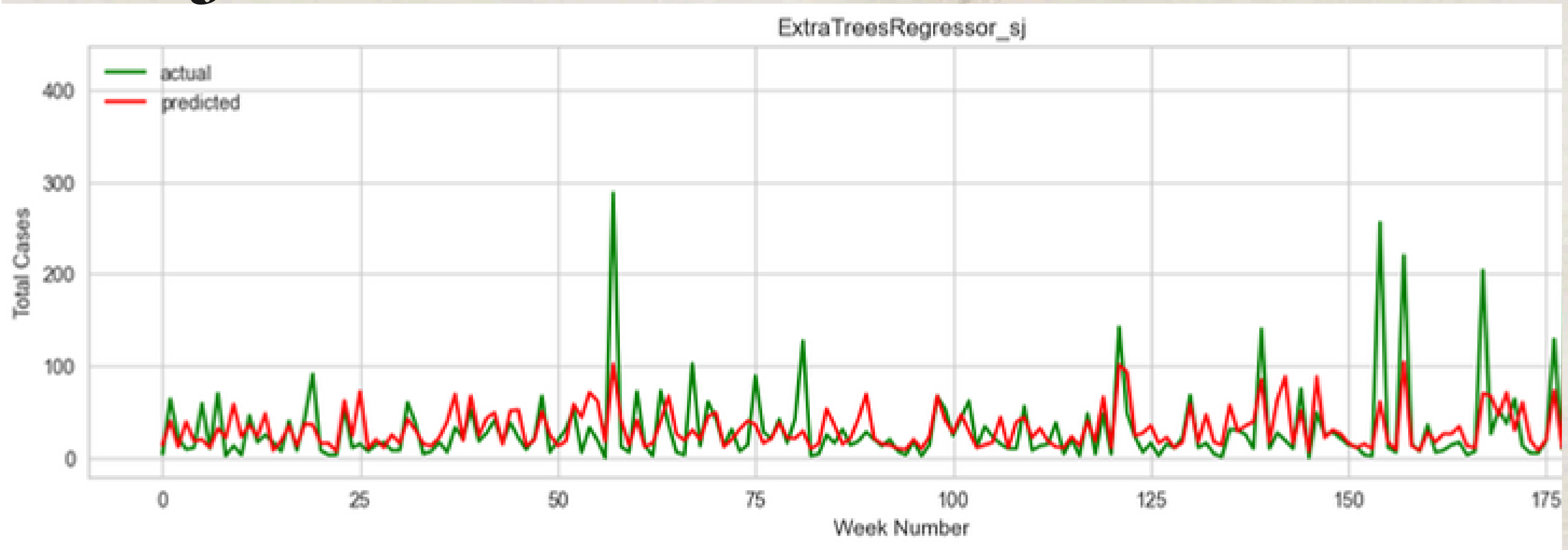
| | Model | MAE | MSE | RMSE | R2 |
|---|---|---|---|---|---|
| **en** | Elastic Net | 6.7681 | 129.8842 | 10.5429 | -0.0253 |
| **omp** | Orthogonal Matching Pursuit | 6.7408 | 129.2654 | 10.5301 | -0.0305 |
| **br** | Bayesian Ridge | 6.7718 | 129.7549 | 10.5521 | -0.0316 |
| **huber** | Huber Regressor | 5.9912 | 138.8084 | 10.7858 | -0.0326 |
| **lasso** | Lasso Regression | 6.7821 | 130.6565 | 10.5735 | -0.0328 |
| **llar** | Lasso Least Angle Regression | 6.7821 | 130.6564 | 10.5735 | -0.0328 |
| **dummy** | Dummy Regressor | 7.0617 | 136.2876 | 10.8050 | -0.0788 |
| **ridge** | Ridge Regression | 6.8976 | 132.1761 | 10.7505 | -0.1074 |
| **lr** | Linear Regression | 6.8753 | 132.8426 | 10.7907 | -0.1214 |
| **et** | Extra Trees Regressor | 6.8037 | 126.9249 | 10.7180 | -0.2536 |

**For Iquitos**

- Shortlisted Models

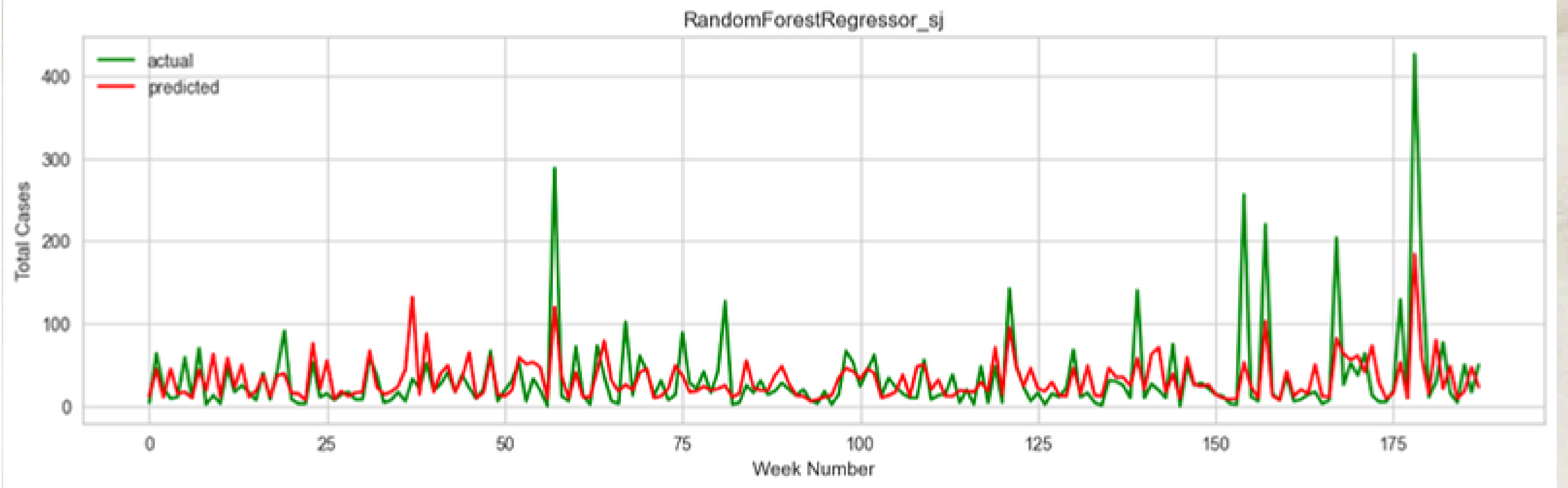
1. HuberRegressor
2. LassoLars
3. ElasticNet
4. Lasso Regression
5. BayesianRidge
6. Ridge
7. OrthogonalMatching
Pursuit
1. LinearRegression
2. ExtraTreesRegressor
3. RandomForestRegressor

# San Juan




**ExtraTreesRegressor:** Builds an ensemble of decision trees during training, but with additional randomness in the feature selection and node splitting process, leading to potentially faster training and improved generalization performance.

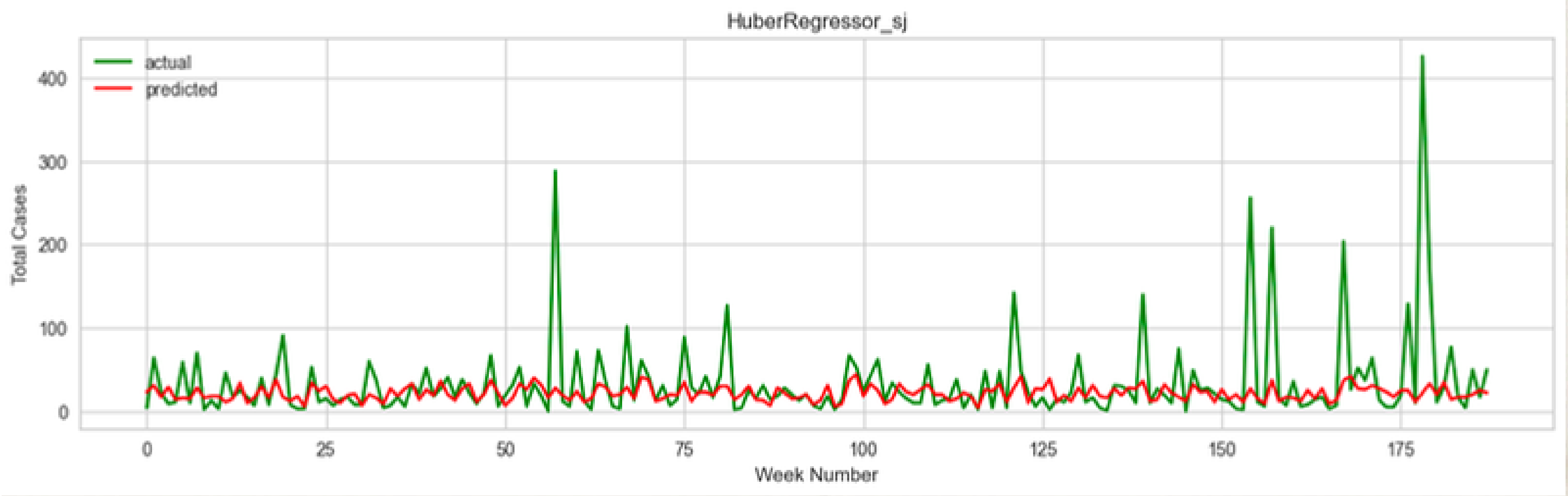
```
Mean Absolute Error (MAE): 21.22340425531915
Mean Squared Error (MSE): 1333.3829787234042
Root Mean Squared Error (RMSE): 36.51551695818374
R-squared (R2 Score): 0.4856616123131029
```




**RandomForestRegressor:** Builds multiple decision trees during training and outputs the average prediction of the individual trees for regression tasks.

Highly effective in handling high-dimensional datasets and can capture complex relationships between input features and the target variable

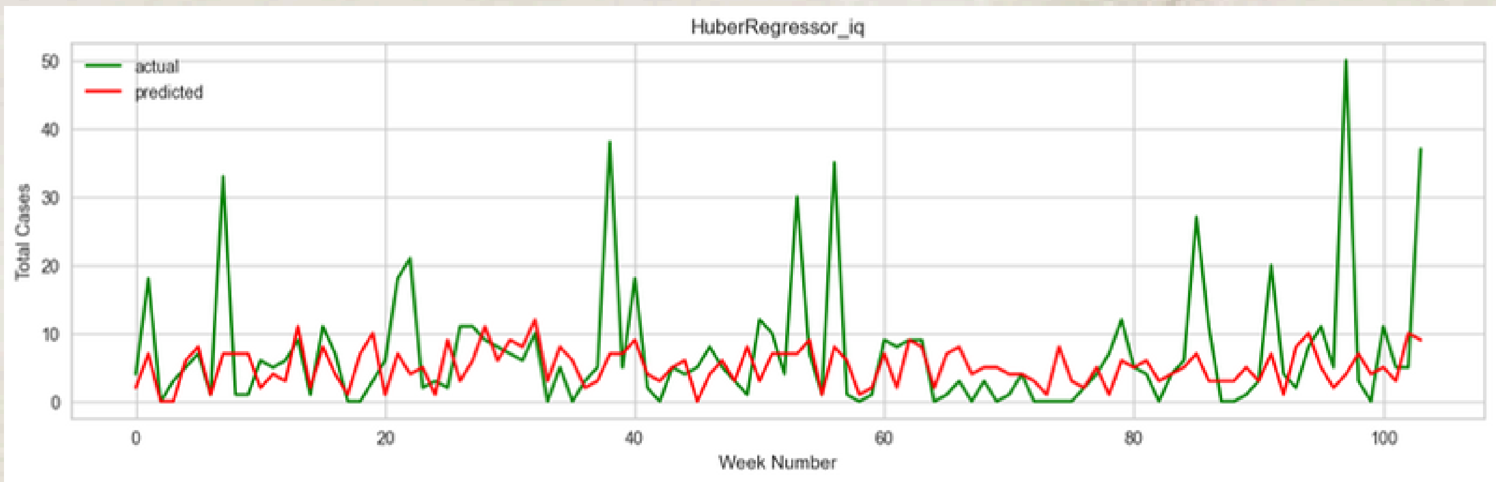
```
Mean Absolute Error (MAE): 21.5
Mean Squared Error (MSE): 1490.3085106382978
Root Mean Squared Error (RMSE): 38.604514122551755
R-squared (R2 Score): 0.4251292473737436
```



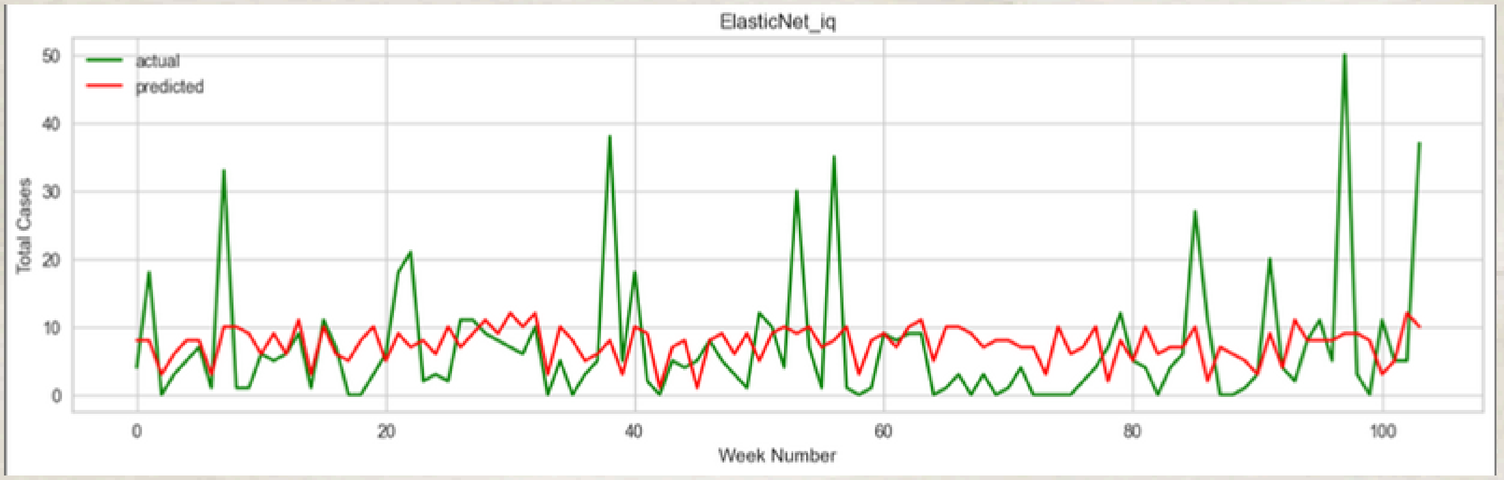

**HuberRegressor:** Combines the robustness of absolute error minimization with the efficiency of squared error minimization. This allows it to handle outliers in the data

```
Mean Absolute Error (MAE): 22.48936170212766
Mean Squared Error (MSE): 2502.7978723404253
Root Mean Squared Error (RMSE): 50.02797089969196
R-squared (R2 Score): 0.034572180006203435
```

# Iquitios

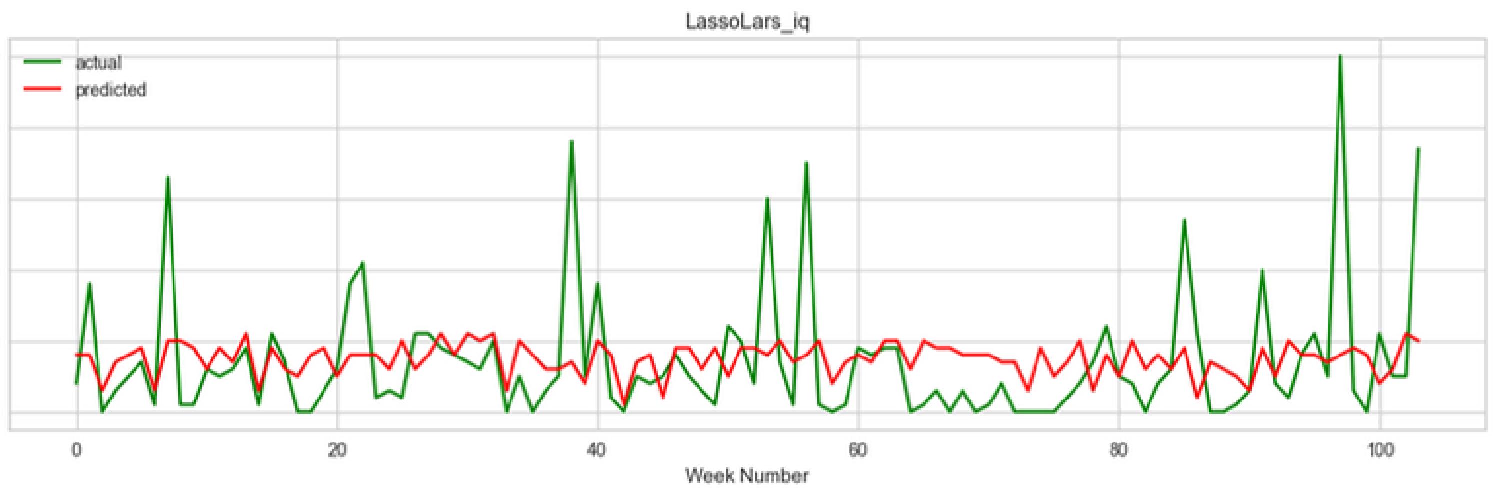



```
Mean Absolute Error (MAE): 5.259615384615385
Mean Squared Error (MSE): 80.0673076923077
Root Mean Squared Error (RMSE): 8.948033733301841
R-squared (R2 Score): 0.041355285282725696
```



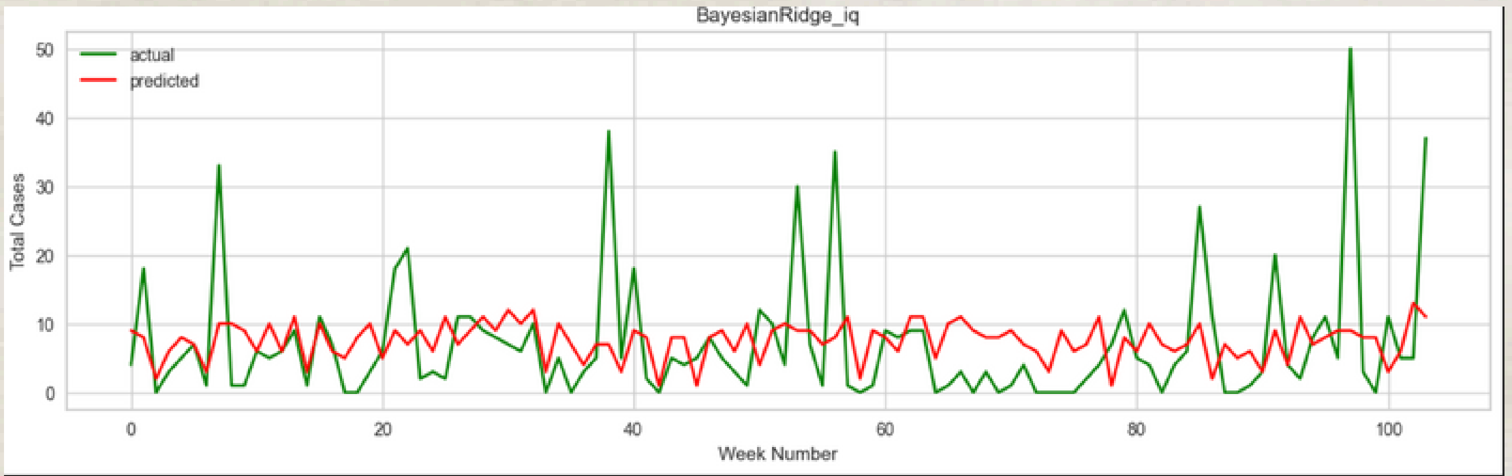

Combines the penalties of both Lasso and Ridgeregression, allowing it to handle multicollinearity and perform feature selection by encouraging sparsity in the coefficients

```
Mean Absolute Error (MAE): 5.990384615384615
Mean Squared Error (MSE): 78.85576923076923
Root Mean Squared Error (RMSE): 8.880077095992423
R-squared (R2 Score): 0.05586101772590768
```




Adds a penalty term to the ordinary least squares method, encouraging sparse feature selection by shrinking the coefficients of less important features towards zero, effectively performing feature selection and regularization simultaneously

```
Mean Absolute Error (MAE): 5.990384615384615
Mean Squared Error (MSE): 79.875
Root Mean Squared Error (RMSE): 8.937281465859739
R-squared (R2 Score): 0.04365778249592911
```




Assumes a Gaussian prior distribution over the coefficients and computes the posterior distribution using the observed data, providing a principled approach to regularization and uncertainty estimation in regression tasks.

```
Mean Absolute Error (MAE): 26.585106382978722
Mean Squared Error (MSE): 2255.095744680851
Root Mean Squared Error (RMSE): 47.487848389675975
R-squared (R2 Score): 0.13012065707542197
```

# Weighted Average of the Selected Models for the final Prediction

## Iquitios

| index | MAE | MSE | RMSE | R2 Score |
|---|---|---|---|---|
| HuberRegressor | 5.2596153846 | 80.0673076923 | 8.9480337333 | 0.0413552853 |
| LassoLars | 5.9903846154 | 79.875 | 8.9372814659 | 0.0436577825 |
| ElasticNet | 5.9903846154 | 78.8557692308 | 8.880077096 | 0.0558610177 |
| Lasso | 6 | 80.0192307692 | 8.9453468781 | 0.0419309096 |
| BayesianRidge | 6.125 | 80.375 | 8.9652105385 | 0.0376712897 |
| Ridge | 6.3942307692 | 85.7019230769 | 9.2575333149 | -0.0261078831 |
| OrthogonalMatchingPursuit | 6.4038461538 | 85.0769230769 | 9.2237152535 | -0.0186247671 |
| LinearRegression | 6.5288461538 | 87.4519230769 | 9.3515732942 | -0.0470606077 |
| ExtraTreesRegressor | 6.7115384615 | 82.5384615385 | 9.0850680536 | 0.0117681961 |
| RandomForestRegressor | 6.7307692308 | 84.9615384615 | 9.2174583515 | -0.0172432688 |

```
final_predictions_iq = (
 0.6 * prediction(iq_std_train.drop('total_cases',axis=1), Y_iq, iq_std_test, 'HuberRegressor') +
 0.3 * prediction(iq_std_train.drop('total_cases',axis=1), Y_iq, iq_std_test, 'ElasticNet') +
 0.05 * prediction(iq_std_train.drop('total_cases',axis=1), Y_iq, iq_std_test, 'LassoLars') +
 0.05  * prediction(iq_std_train.drop('total_cases',axis=1), Y_iq, iq_std_test, 'BayesianRidge')).astype(int)
```

- 4 Models performed better than 3, based on test submission (MAE 25.5505 vs  25.3101)
  - HuberRegression: 60%
  - ElasticNet: 30%
  - LassoLars: 5%
  - BasyesianRidge: 5%

## San Juan

| index | MAE | MSE | RMSE | R2 Score |
|---|---|---|---|---|
| ExtraTreesRegressor | 21.2234042553 | 1333.3829787234 | 36.5155169582 | 0.4856616123 |
| RandomForestRegressor | 21.5 | 1490.3085106383 | 38.6045141226 | 0.4251292474 |
| HuberRegressor | 22.4893617021 | 2502.7978723404 | 50.0279708997 | 0.03457218 |
| Lasso | 26.2925531915 | 2237.1968085106 | 47.2990148789 | 0.1370249825 |
| LassoLars | 26.2978723404 | 2237.2446808511 | 47.2995209368 | 0.1370065162 |
| BayesianRidge | 26.585106383 | 2255.0957446809 | 47.4878483897 | 0.1301206571 |
| ElasticNet | 26.6117021277 | 2313.7925531915 | 48.1018976049 | 0.1074789837 |
| OrthogonalMatchingPursuit | 27.2180851064 | 2310.3457446809 | 48.066056055 | 0.1088085537 |
| Ridge | 27.3670212766 | 2217.6436170213 | 47.0918635968 | 0.1445674194 |
| LinearRegression | 27.5212765957 | 2224.414893617 | 47.1637031372 | 0.141955471 |

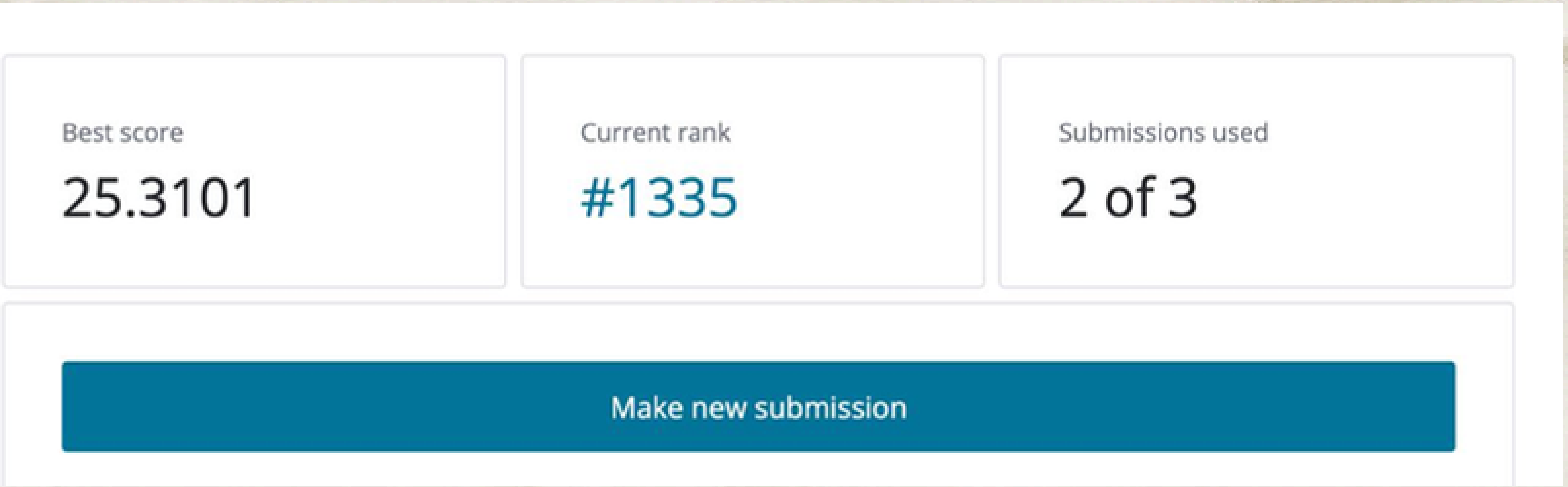
```
final_predictions_sj = (
    0.6 * prediction(sj_std_train.drop('total_cases',axis=1), Y_sj, sj_std_test,'ExtraTreesRegressor') +
    0.3 *  prediction(sj_std_train.drop('total_cases',axis=1), Y_sj, sj_std_test,'RandomForestRegressor',sj_rf_params)
    0.1 *  prediction(sj_std_train.drop('total_cases',axis=1), Y_sj, sj_std_test,'HuberRegressor')).astype(int)
```

- Top 3 models, shows a good trend in test split and backed by literature survey
  - Extra Tree Regressor: 60%
  - Random Forest Regressor: 3 0%
  - Huber Regressor: 10%

# 06

# Results

# Ranking in DrivenData DengAI




Total Participants: 14,990
Total Teams: 5994
Our Rank: 1335
Top 20% of Teams

# Deployability at Plaksha

- Implement this on the Plaksha Campus, as a preventive mechanism for predicting and preparing for vector-borne diseases like Dengue.
- It can also be used by Plaksha Health Department to better prepare and acquire any logistics in case of an outbreak in atleast 3 weeks in advance
- Given the current temperature range in Chandigarh (33 - 38 degrees Celsius), it's apparent that the climate differs significantly from coastal regions where our model was originally trained. In the dataset, we observed a strong correlation between temperatures ranging from 25 to 30 degrees Celsius and reported dengue cases. However, with the higher temperatures experienced in Chandigarh, this correlation may undergo changes

# Challenges

- As the deployment scales up to cover larger geographic areas or multiple regions, integrating heterogeneous data sources from various sources becomes challenging. Standardizing data formats, ensuring interoperability between different systems, and addressing data quality issues across diverse datasets are some challenges

- Handling large volumes of data and complex computational tasks associated with model training, validation, and inference requires significant computational resources.

- Models trained on data from specific regions may not generalize well to new geographic areas or populations with different environmental conditions, demographics, and healthcare infrastructure.

- Limited resources, including financial, human, and infrastructural resources, may constrain the scalability of the dengue prediction solution, particularly in resource-constrained settings or low-income regions.

# References

[1] N. A. M. Salim et al., "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques," Sci. Rep., vol. 11, no. 1, pp. 1–9, 2021.
[2] P. Méndez-Lázaro, F. Muller-Karger, D. Otis, M. McCarthy, and M. Peña-Orellana, "Assessing climate variability effects on dengue incidence in San Juan, Puerto Rico," Int. J. Environ. Res. Public Health, vol. 11, no. 9, pp. 9409–9428, 2014.
[3] "Historic data (2010-2023)," Cdc.gov, 09-Feb-2024. [Online]. Available: https://www.cdc.gov/dengue/statistics-maps/historic-data.html. [Accessed: 16-Mar-2024].
[4] Indjst.org. [Online]. Available: https://indjst.org/articles/using-public-open-data-to-predict-dengue-epidemic-assessment-of-weather-variability-population-density-and-land-use-as-predictor-variables-for-dengue-outbreak-prediction-using-support-vector-machine. [Accessed: 16-Mar-2024].
[5] B. Narapusetty, T. DelSole, and M. K. Tippett, "Optimal estimation of the climatological mean," J. Clim., vol. 22, no. 18, pp. 4845–4859, 2009.
[6] P. Guo et al., "Developing a dengue forecast model using machine learning: A case study in China," PLoS Negl. Trop. Dis., vol. 11, no. 10, p. e0005973, 2017.
[7] B. M. Althouse, Y. Y. Ng, and D. A. T. Cummings, "Prediction of dengue incidence using search query surveillance," PLoS Negl. Trop. Dis., vol. 5, no. 8, p. e1258, 2011.
[8] R. Tuladhar et al., "Effect of meteorological factors on the seasonal prevalence of dengue vectors in upland hilly and lowland Terai regions of Nepal," Parasit. Vectors, vol. 12, no. 1, 2019.

# Thanks!