Forecasting Dengue Fever Epidemics: San Juan and Iquitos

1st Samiksha Sharma
University Of Nottingham
Nottingham, UK
psxss39@nottingham.ac.uk

2nd Vaishali Gupta
Computer Science
University Of Nottingham
Nottingham, UK
psxvg2@nottingham.ac.uk

3rd Shweta Bhati
Computer Science
University Of Nottingham
Nottingham, UK
psxsb14@nottingham.ac.uk

Abstract—**Dengue fever, an illness transmitted by Aedes mosquitoes and caused by RNA viruses, poses a significant health threat in tropical regions. This paper aims to investigate the transmission dynamics of dengue, particularly its correlation with fluctuating climate factors like temperature and precipitation. Variations in climate can influence the spread of dengue, potentially impacting public health. The primary goal of this study is to predict the expansion of dengue fever in two urban centers, San Juan and Iquitos, utilizing historical data covering several years. In this report, we utilize regression and statistical models to predict the occurrence of Dengue cases. The study evaluates model performance and proposes potential enhancements, highlighting the significance of data-driven approaches in tackling public health challenges.**

I. Introduction

Dengue fever, an acute systemic viral illness, has established itself globally in both endemic and epidemic transmission cycles. Mosquitoes transmit the disease and represent a significant public health challenge in tropical regions. Dengue virus infection in humans can range from mild fever to potentially fatal dengue shock syndrome, although it often goes unnoticed[4]. The World Health Organization has categorized it into severe and uncomplicated types. Severe dengue can have life-threatening consequences, including organ impairment, bleeding, or plasma leakage [1]. To date, four virus stereotypes have been identified: DENV-1, DENV-2, DENV-3, and DENV-4 [2]. The Dengue virus enters the host organism through the skin when an infected mosquito bites. The body's immune system responds with humeral, cellular, and innate defenses to fight the illness. Interestingly, the more severe symptoms occur when the virus is quickly cleared from the host organism, not when the viral load is high[3].

Dengue fever has a well-documented historical spread; however, accurately assessing its implications on health across tropical and subtropical regions remains a challenge owing to the factors related to limited diagnostic capabilities and inadequate disease surveillance systems. Despite the current efforts to control dengue fever, it is clear that targeting the Aedes mosquitoes - responsible for transmitting the virus - through chemical and biological methods, and managing breeding sites to reduce mosquito populations, have not been enough to effectively stem the growing number of outbreaks. The virus continues to expand into new areas, and we need to explore new control strategies to put a stop to this deadly disease[5]. The present study undertakes an in-depth analysis of the transmission dynamics of dengue fever in two key locations, namely San Juan and Iquitos. The study leverages a comprehensive dataset that combines environmental variables and historical dengue case counts intending to develop predictive models. Our objective is to offer insights into the factors influencing disease transmission and to enable accurate forecasting of outbreak scenarios. Our study employs rigorous statistical analysis and advanced modeling techniques to contribute to the ongoing efforts to combat dengue fever. The work presented here aims to provide actionable insights and tools that can be used to enhance the effectiveness of public health decision-making.

II. Related Works

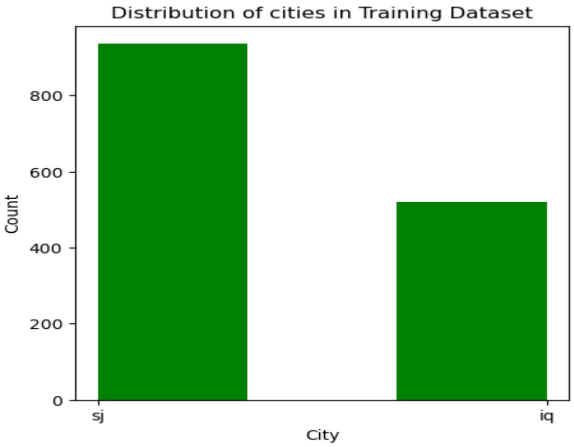
Before diving deep into the report, let's first take a look at some of the significant work that has been done in this area by researchers. The author of [6] compared the ability of ML, regression, and time-series-based modeling approaches to forecasting dengue case counts and outbreaks. Random Forest excelled in short-term dengue forecasts and ARIMA for the long term. RF-UFA enhanced outbreak prediction. ML models improve dengue early warning systems. In [7], the author used linear correlation to select the best model and a weighted Poisson GLM for modeling. This approach improved data fitting, robustness, and interpretability. The author of [8] used Raman spectroscopy for medical diagnoses, including cancer and infectious diseases. classified serum samples of healthy people and dengue patients with RF, achieving 91 percent accuracy. The same author published another paper using SVM but had lower accuracy (85 percent) than RF had achieved. In [9], the author proposed a Negative Binomial model taking Mean, Average, Maximum Temperature, and monthly Cumulative Rainfall values as the dataset and observed that the Temperature and Rainfall specifically are linked with the incidence of Dengue fever. In [10], the author used multiple classifiers for a predictive model for the outbreak of Dengue. The various classifiers used are Decision Trees, Rough Set Classifiers, Associative Classifiers, and Nave Bayes. The combination of all the classifiers gives more accuracy as compared to the accuracy of a single classifier. In [11], the author described the model for the prediction of Dengue outbreak using a vector correction method occupying Relative Humidity and Temperature only. In [12], the author presents a model for the Dengue outbreak prediction that associates the least square support vector machine using Dengue cases and measure of rainfall level.

III. Data Handling

A. Data Set Information

For our research project, we have obtained our dataset from the competition run by the Data-driven organization. The data contains different environmental factors collected by various U.S. Federal Government agencies—from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce[13].

Our dataset includes several meteorological and environmental variables that are closely linked to the transmission dynamics of diseases like dengue fever. These variables comprises of temperature, precipitation, and vegetation indices for San Juan and Iquitos. Moreover, our data includes time series features such as year, ”week_of_year”, and ”week_start_date”. To improve the accuracy of our model we have separated our dataset into two based on city as both the cities have different environmental factors affecting the temperature. Going forward, we will be cleaning, exploring, and modeling them separately through the entire project pipeline.



H

Fig. 1

B. Data Cleaning

As our dataset is raw and may contain irregularities such as missing values, inconsistencies, etc., it can affect our model's performance. To address this issue, we need to check our dataset for any irregularities.

Upon checking, we found that the dataset does contain irregularities in the form of missing values marked as ”Null” or ”NAN”. For San Juan, the maximum number of null values is 191 in the NDVI feature, while for Iquitos, it is a maximum of 37 in ”station_avg_temp_ c” and ”station_ diur_temp_rng_c”. Overall, San Juan dataset has 936 rows and Iquitos has 520 rows. Since the maximum number of null values in a column of this dataset is very large compared to the entire size of the dataset, dropping these rows containing null values is not a feasible solution. Therefore, we have handled these missing values by using the simple imputer mean strategy to fill these values.

C. Scaling

Scaling input features is essential in predictive modeling to normalize their impact. Our dataset shows NDVI features already fall within the -1 to 1 range with a mean close to zero, negating the need for scaling. Conversely, weather features exhibit diverse scales; precipitation may range from 0-390, while ”reanalysis_ tdtr_k ranges from 1-5. Thus, we opted to scale only weather features in city datasets, enhancing our predictive model accuracy and reliability further down the pipeline.

D. Exploratory Data Analysis

Exploratory data analysis is performed to analyze attributes and summarize their characteristics using statistical techniques to discover useful patterns and graphical representations[15]. Our exploratory data analysis (EDA) focuses on understanding the relationship between environmental variables and dengue transmission patterns. We have examined the correlation of all environmental features with one another. Visualized the distribution of key features, such as vegetation index, and weather features such as temperature, and rainfall over time and across different regions within both cities. Furthermore, we also examined the correlations between all vegetation indices and environmental factors to identify potential predictors for our models.

Exploring the correlation of Features: We have constructed a simple heat map for all the features of both cities which illustrates: the satellite imagery scores of the vegetation growing in both cities exhibit some interdependence with one another. There is a high correlation within the North quadrant and also within the South quadrant in San Juan. The correlation in Iquitos is even stronger both within and between the north and South quadrants.

The weather features also have a few columns that have a high correlation for example in Iquitos columns like ”reanalysis_air_temp_k” and ”reanalysis_avg_temp_k” have a very high correlation with one another and in San Juan same columns plus some others show a high correlation.

These observations can be seen in Figure 2 and Figure 3.

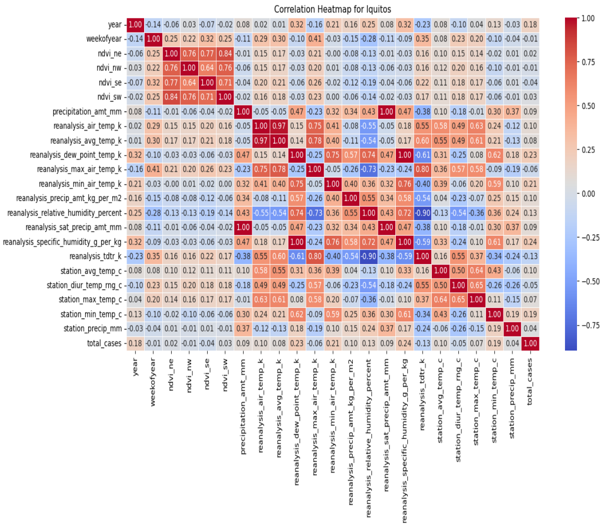


Fig. 2

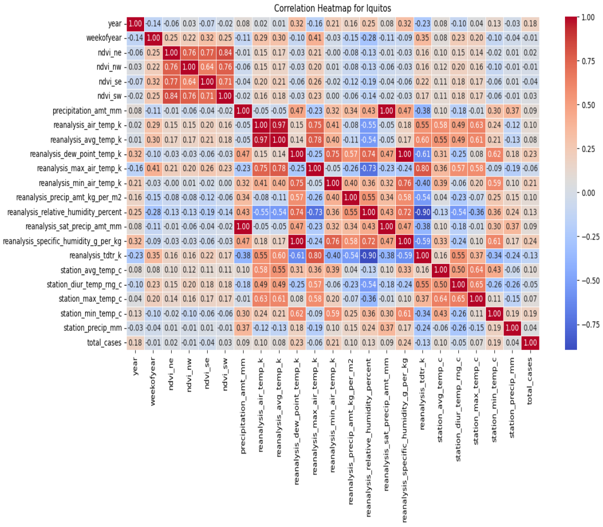


Fig. 3

Exploring NDVI features We observed the distribution of NDVI features over time by averaging their scores over the 10 years for each week and plotting them on time graphs. This confirmed our previous observations about them from the correlation heat maps. we can see these in Figure 4 and Figure 5 for both the cities.

In Figure 4, San Juan's NDVI scores depict consistently lower values in the Northeast and Northwest regions compared to the Southeast and Southwest quadrants. Conversely, Figure 5 illustrates a more uniform NDVI score distribution across all quadrants in Iquitos. Another significant observation is the fluctuation in vegetation levels in San Juan, with minimal variations, whereas Iquitos exhibits spikes in vegetation levels during the second half of each year.

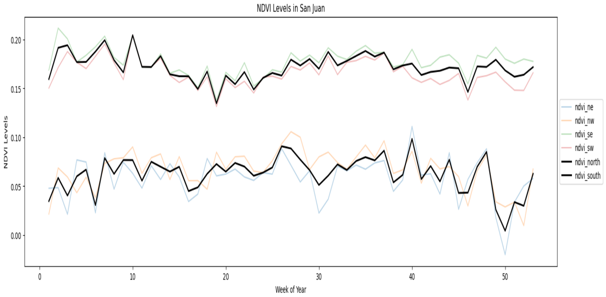


Fig. 4

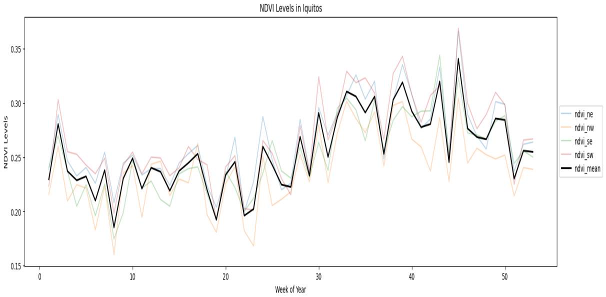


Fig. 5

Exploring Weather features We have constructed a simple line graph by averaging weather features over 10 years for each week for both cities. In Figure 6 for San Juan also, we observed the same thing as we did in the correlation map, that is the correlation between the five features, 'reanalysis_air_temp_k', 'reanalysis_avg_temp_k', 'reanalysis_dew_point_temp_k', 'reanalysis_max_air_temp_k', 'reanalysis_min_air_temp_k' is very high whereas, Figure 7 of Iquitos shows that only 2 features, 'reanalysis_air_temp_k', 'reanalysis_avg_temp_k' have such a high correlation tendency. From the list of weather features, we can see that we are using overlapping features to measure the form of temperature, humidity, and rainfall. This in turn leads to the trend variation in San Juan and Iquitos as both the cities are in a different hemisphere.

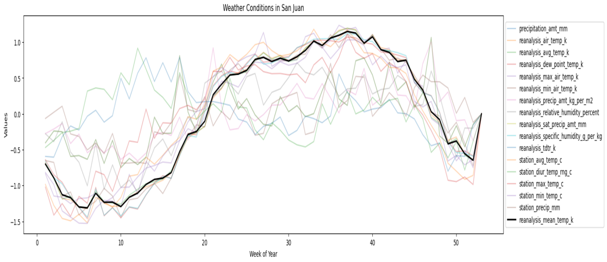


Fig. 6

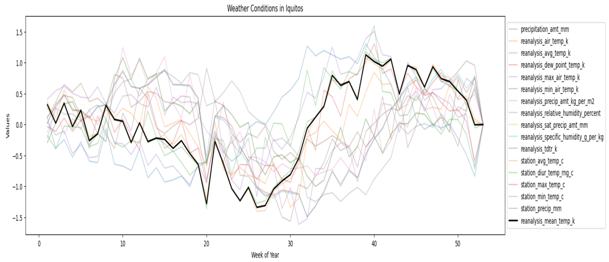


Fig. 7

Exploring the target Now that we have measured the importance of each feature against each other and with time let's also visualize our target variable i.e. total cases against all features and against time in both cities.

Correlation with features From Figures 8 and 9, we can see that none of the features have a strong correlation with our target variable, but since there is a positive correlation in most of them, we will do predictive modeling using the regression model. We can also use the three most highly correlated features in our statistical modeling.

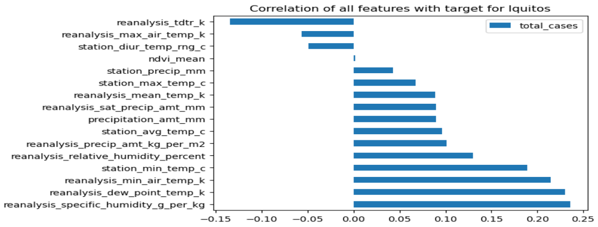


Fig. 8

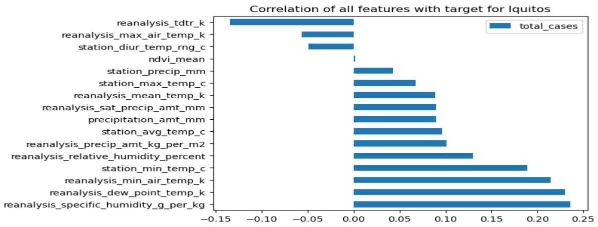


Fig. 9

From Figures 10 and 11, we can observe that there are outbreaks in both Iquitos and San Juan toward the end of each year. The increases in cases and outbreaks tend to happen in weeks 35–45 in San Juan and in weeks 45–50 in Iquitos. It can also be seen that most outbreaks occurred during the end of the year 2004 in both cities. Another significant year of dengue outbreaks before this, but not as high as this, was 1988. The dengue outbreak was prevalent during the middle of the year in San Juan and the end of the year in Iquitos.

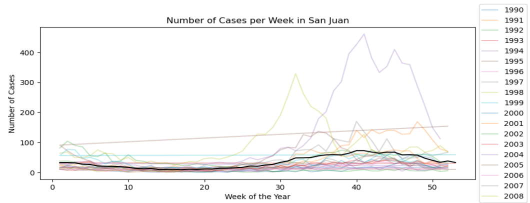


Fig. 10

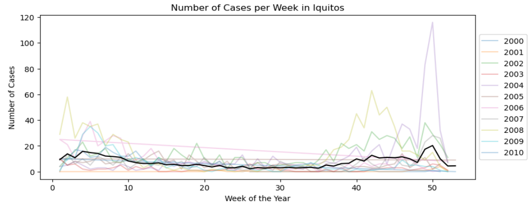


Fig. 11

E. Feature Engineering

Feature engineering is a process that enhances the accuracy of dengue fever prediction models by creating new features from environmental data. In this process, NDVI features are consolidated for San Juan, while each city's reanalysis temperature variables are averaged. Redundant features are eliminated from the datasets, which streamlines them. Effective feature engineering improves the accuracy of the given model and provides deeper insights into the dynamics of dengue fever. We figured we could combine these Normalized Difference Vegetation Index (NVDI) scores into fewer dimensions, e.g. In San Juan, the north quadrant can be reduced to a single dimension called ”ndvi_north,” and similarly, the south quadrant can be reduced to another dimension called ”ndvi_south.”. On the other hand, in Iquitos, we can combine all four ”NDVI” features into one dimension called ”ndvi_mean”.Similarly, weather features can be reduced to fewer dimensions by combining those highly correlated into a single feature called ”reanalysis_mean_temp_k” for both cities.

IV. Modelling

A. Machine Learning Models

In our analysis, we have developed different predictive models based on time features, meteorological features, and weather features. Figure 12, reflects the general framework, used in this study, for developing a predictive ML (i.e., RF, linear regression, Gradient boost regression, etc) and time-based models (i.e., Negative binomial, Naive Bayesian) using historical data as input. Since regression is one of the most valuable tools for creating predictive models, we have used it for our analysis.

Hyper-parameter Tuning We have performed hyperparameter adjustments before beginning the model fitting phase. Hyperparameter tuning helps to find the best possible set of parameters for a model, improving accuracy and generalization

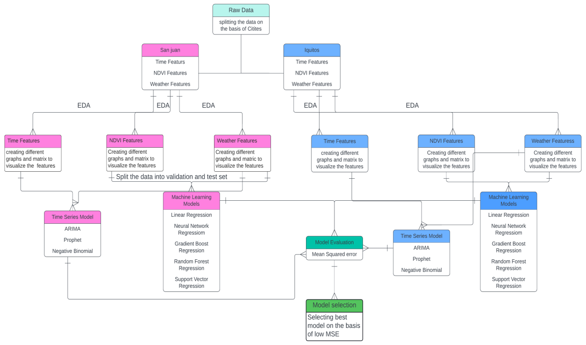

Raw Data
splitting the data on the basis of Citites
San juan
Time Featurs
NDVI Features
Weather Features
Iquitos
Time Features
NDVI Features
Weather Features
EDA
EDA
EDA
EDA
EDA
Time Features
creating different graphs and matrix to visualize the features
NDVI Features
Creating different graphs and matrix to visualize the features
Weather Features
Creating different graphs and matric to visualize the features
Time Features
creating different graphs and matrix to visualize the features
NDVI Features
Creating different graphs and matrix to visualize the features
Weather Featuress
Creating different graphs and matrix to visualize the features
Split the data into validation and test set
Time Series Model
ARIMA
Prophet
Negative Binomial
Machine Learning Models
Linear Regression
Neural Network Regression
Gradient Boost Regression
Random Forest Regression
Support Vector Regression
Model Evaluation
Mean Squared error
Time Series Model
ARIMA
Prophet
Negative Binomial
Machine Learning Models
Linear Regression
Neural Network Regression
Gradient Boost Regression
Random Forest Regression
Support Vector Regression
Model selection
Selecting best model on the basis of low MSE

Fig. 12

while reducing overfitting. We have used grid search on our models to do the hyper-parameter tuning. Grid search exhaustively searches through a predefined grid of hyper-parameter values, evaluating model performance for each combination using cross-validation to find the optimal hyper-parameters[16].In this case, our cross-validation has been done on both cities' datasets, each with 5 K-fold splits.

B. Models Tested

Below are a few of the Regression models that we have tested on both the city's datasets.

Linear Regression Model: Linear regression is considered one of the fundamental yet robust statistical techniques for predictive modeling. Hence, we start by initializing a linear regression model (”model_1”) as the foundation of our predictive modeling framework. Here, we are trying to model the relationship between the total dengue cases and one or more independent variables which in our case are weather and meteorological. Since linear regression does not require hyperparameter tuning as other algorithms do, we have defined an empty parameter grid for standardized model evaluation and comparison.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Random Forest Regressor: To effectively capture complex relationships in the data and cover the nonlinear pattern of dengue transmission in both cities, we have used a Random Forest regressor (model_2) on the training dataset of two cities. This model combines multiple decision trees to create a robust predictive model. It also measures the importance of different features, such as weather, NDVI, and time, on the total number of cases.

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$$

To ensure the optimal performance and generalization ability of the model, we have specified a range of values for the number of estimators in our hyperparameter grid (param_grid_2), which will be evaluated.

Gradient Boosting Regression: We have used a powerful ensemble learning technique to analyze our training data for two cities. This technique involves building an ensemble of weak learners (usually decision trees) successively, with each subsequent model focusing on the residuals of the previous models. To capture the complex interactions and nonlinear relationships in our data, we used gradient boost (model_3). We also utilized two hyperparameters, namely the n estimator and the learning rate. The n estimator value was set to 50, 100, or 150, and it controlled the length of the decision trees. The learning rate value was set to 0.01, 0.1, or 0.2, and it helped to shrink the influence of each new tree on the overall model.

$$\hat{y} = F(x) + \sum_{i=1}^{N} \gamma_i h_i(x)$$

Support Vector Regression: SVR, is a potent supervised learning algorithm adept at identifying a hyperplane within a multidimensional space for accurate classification and prediction of the data. It seeks to widen the margin between predicted and actual values while minimizing error. To achieve this (model_4), we have utilized an RBF kernel, a well-known function in machine learning. Moreover, through grid search, we fine-tuned hyperparameters like regularization parameter (C) and kernel coefficient (gamma), thereby enhancing model efficacy.

$$y = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b$$

Neural Network: We have customized neural networks for San Juan and Iquitos datasets to optimize the accuracy of dengue fever. KerasRegressor wrappers were employed for seamless integration with scikit-learn. Hyperparameters grids were defined for tuning, focusing on variations in neurons and optimizers. These models enable fine-tuning to account for the unique characteristics of each city's data, enhancing our ability to predict and understand dengue fever outbreaks.

$$\hat{y} = f(W^{(2)} \cdot f(W^{(1)} \cdot x + b^{(1)}) + b^{(2)})$$

1) Time Series Model: In the given dataset, we have time-based features such as week_of_year, year, and, week_start_date. However, we will only use week_start_date as the main time feature for modeling and forecasting. We have applied different time series techniques to capture the seasonal patterns and trends in our dataset which further aid in modeling and forecasting dengue fever.

ARIMA Model: Dengue fever outbreaks exhibit seasonal cycles, and ARIMA models are proficient in capturing the patterns and making predictions based on historical data. In our model, we have created ”Best_Param_ARIMA” function which searches for the optimal parameters (p, d, q) for an

ARIMA model by iterating through a predefined range [which is (0,2of values for each parameter. It evaluates each combination using the Akaike Information Criterion (AIC) to measure the model's goodness of fit. The combination with the lowest AIC score is considered the best. This process ensures the selection of the most suitable ARIMA parameters for accurate time series forecasting.

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$$

Negative Binomial Model: The negative binomial relaxes the assumption of equal population mean and variance, thereby allowing a wider range of possible models. We have created a ”Get_Stats_NegBinomialModel” function which models the relationship between dengue fever cases, time, and environmental factors such as temperature. This is done by creating a model formula in which, we give all the significant features used to create the model and fit the data. In our case the model formulae for San Juan have features ’total_cases’, ’reanalysis_specific_humidity_g_per_kg’, ’station_avg_temp_c’, ’reanalysis_mean_temp_k’, and the formulae for Iquitos has features ’total_cases’, ’reanalysis_specific_humidity_g_per_kg’, ’reanalysis_dew_point_temp_k’, ’reanalysis_min_air_temp_k’ iterates over a range of alpha values to find the optimal regularization parameter, ensuring the best model fit.

$$y_i \sim \text{NegBinomial}(\mu_i, \phi)$$

Prophet Model: It is a powerful tool for time series forecasting, used to predict the incidence of dengue fever cases. It pre-processes the data by taking ”week_start_date” as a date time column and ”tota_cases” as the target column and renaming them to ’ds’ and ’y’ respective, adhering to Prophet's input requirements. Subsequently, the function fits the Prophet model to the data and generates forecasts for future periods.

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

V. Model Evaluation

Once all regression and time series forecasting models are trained for both cities, the performance of the models is evaluated using mean squared error.

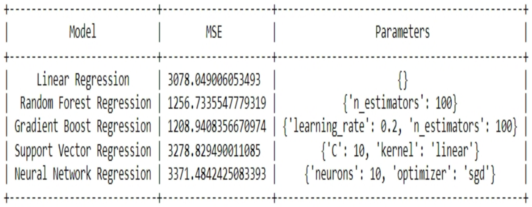
Mean Squared Error:

Mean Squared Error (MSE) is a measure of the average squared difference between the estimated values and the actual values. It is commonly used to evaluate the performance of regression models. Lower MSE values indicate better models. MSE is evaluated by the Equation[15]:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

In our analysis for predicting dengue cases in Iquitos and San Juan, we first assessed various machine learning regression models mentioned in the earlier section, by fitting them on the training data with their best parameters, then predicting and assessing them using their MSE score.

Figures 13 and 15, present the Mean Squared Error (MSE) values for all these regression models and their tuned hyperparameters for San Juan and Iquitos respectively. we also represented these MSE values on a graph to compare them visually. On analyzing the below figure, we can conclude that Gradient boost regression(with hyperparameter learning rate = 0.2 and n_estimator = 100) gives the least error with its prediction on San Juan data, and Random forest regression(with hyperparameter n_estimator = 300) gives the least error for the same on Iquitos data. We have finalized these two as the best regression models, made predictions using them on the test splits of our train datasets for the two cities, and visually compared them with the actual values of the test splits through graphs in Figures 14 and 16.   We also assessed various time series models mentioned in the earlier section by fitting them on the training data, forecasting, and, then assessing them using their MSE score. Figures 17 and 18, present the mean squared error values for all these time series



```
+---------------------------+-------------------+------------------------------------------+
|           Model           |        MSE        |                Parameters                |
+---------------------------+-------------------+------------------------------------------+
|     Linear Regression     | 3078.049006053493 |                    {}                    |
|  Random Forest Regression | 1256.7335547779319|           {'n_estimators': 100}          |
| Gradient Boost Regression | 1208.9408356670974| {'learning_rate': 0.2, 'n_estimators': 100} |
| Support Vector Regression | 3278.829490011085 |        {'C': 10, 'kernel': 'linear'}     |
| Neural Network Regression | 3371.4842425083393|      {'neurons': 10, 'optimizer': 'sgd'} |
+---------------------------+-------------------+------------------------------------------+
```

Fig. 13

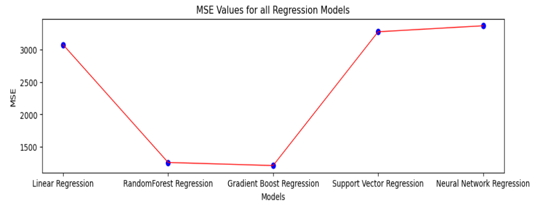
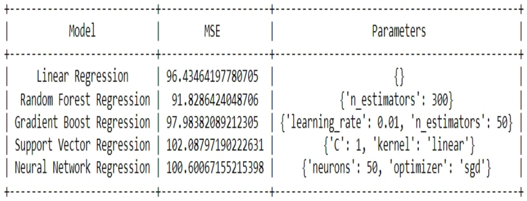


Fig. 14



```
+---------------------------+-------------------+------------------------------------------+
|           Model           |        MSE        |                Parameters                |
+---------------------------+-------------------+------------------------------------------+
|     Linear Regression     | 96.43464197780705 |                    {}                    |
|  Random Forest Regression |  91.8286424048706 |           {'n_estimators': 300}          |
| Gradient Boost Regression | 97.98382089212305 | {'learning_rate': 0.01, 'n_estimators': 50} |
| Support Vector Regression | 102.08797190222631|        {'C': 1, 'kernel': 'linear'}      |
| Neural Network Regression | 100.60067155215398|      {'neurons': 50, 'optimizer': 'sgd'} |
+---------------------------+-------------------+------------------------------------------+
```

Fig. 15

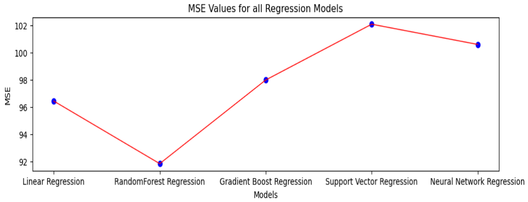
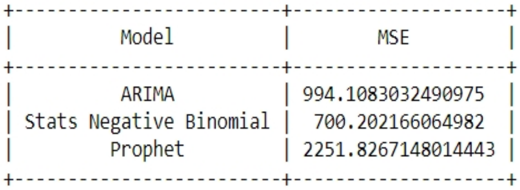


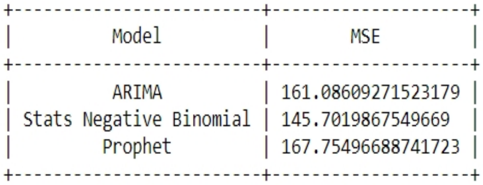
Fig. 16

models with their visual time representation on a graph for comparison. The stats negative binomial is the best statistical model for both the San Juan and Iquitos datasets.



```
+--------------------------+--------------------+
|          Model           |        MSE         |
+--------------------------+--------------------+
|          ARIMA           | 994.1083032490975  |
| Stats Negative Binomial  |  700.202166064982  |
|         Prophet          | 2251.8267148014443 |
+--------------------------+--------------------+
```

Fig. 17



```
+--------------------------+-------------------+
|          Model           |        MSE        |
+--------------------------+-------------------+
|          ARIMA           | 161.08609271523179 |
| Stats Negative Binomial  | 145.7019867549669  |
|         Prophet          | 167.75496688741723 |
+--------------------------+-------------------+
```

Fig. 18

Figures 19 and 20, shows we have again represented the predictions from this model(Negative Binomial) versus the actual values of the test split of our training data on a graph for visual comparison.

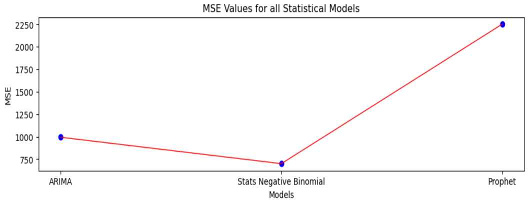


Fig. 19

VI. Model Prediction On Final Test data

Since our test data has the same features as our train data, it is required to split our final test into two: one for San Juan City and another for Iquitos. Both these datasets now need to be pre-processed similar to our training data. It needs to be cleaned by imputing its missing values using the same

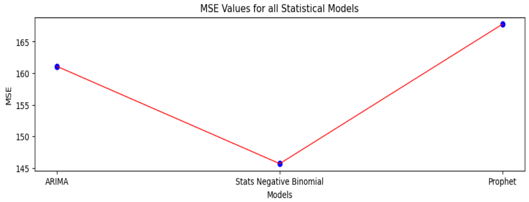


Fig. 20

strategy as used for the training data (Simple imputer with mean strategy). Then it needs to be scaled for all its weather features and then go through the same feature engineering as was done for our training data sets.

Once we preprocess our final test data this way, then we can make predictions on it using the selected best models for each city(for both regression and time series). These predictions, along with their respective ’year’ and ’weekofyear’, were then merged into one and then converted into a CSV output format. Hence, we finally produced two output files, one with predictions using the best regression model (gradient boost and random forest) and one with predictions using the best statistical model (negative binomial). We also plotted this prediction on a graph, as shown in Figures 21, 22, 23, and 24.

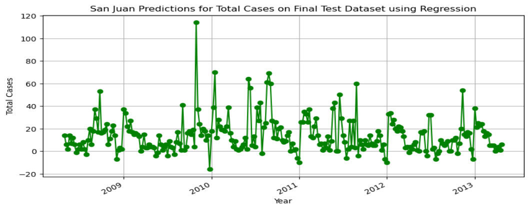


Fig. 21

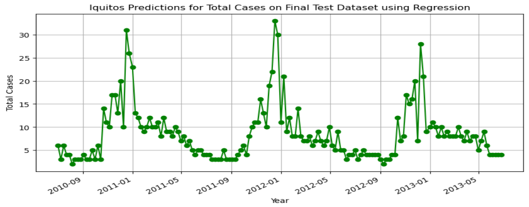


Fig. 22

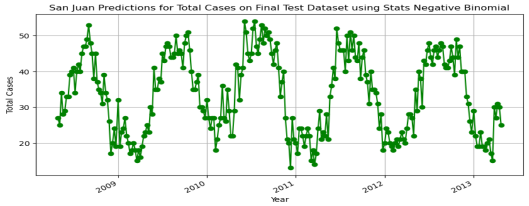


Fig. 23

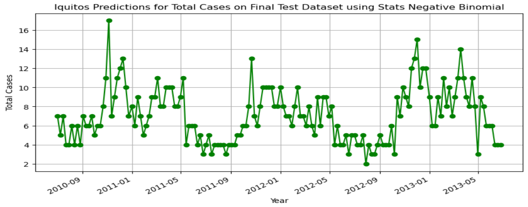


Fig. 24

As per both, our regression and statistical model San Juan has more possibility of dengue outbreak with more number of cases than Iquitos in the coming year until 2013.

VII. Conclusion

The study employs regression and statistical models on comprehensive datasets to forecast Dengue fever outbreaks in San Juan and Iquitos, emphasizing data-driven approaches in public health. Despite control efforts, Dengue persists, necessitating new strategies. The model evaluation identifies Gradient Boosting Regression and Random Forest Regression as the most effective for prediction, alongside statistical models like the Negative Binomial Model. San Juan shows higher outbreak potential than Iquitos until 2013. These findings underscore the significance of advanced analytical techniques, including both machine learning and statistical approaches, in understanding disease dynamics and guiding public health decisions effectively.

References

[1] Hasan S, Jamdar SF, Alalowi M, Al Ageel Al Beaiji SM. Dengue virus: A global human threat: Review of literature. J Int Soc Prev Community Dent. 2016 Jan-Feb;6(1):1-6. doi: 10.4103/2231-0762.175416. PMID: 27011925; PMCID: PMC4784057.

[2] Martina B, Koraka P, Osterhaus A. Dengue virus pathogenesis: an integrated view. 2009. https://doi.org/10.1128/CMR.00035-09.

[3] Whitehorn J, Simmons CP. The pathogenesis of dengue. Vaccine. 2011;29:7221–8. [PubMed] [Google Scholar]

[4] World Health Organization. Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control. WHO/HTM/NTD/DEN/2009.1 (World Health Organization, 2009)

[5] Bhatt, S., Gething, P., Brady, O. et al. The global distribution and burden of dengue. Nature 496, 504–507 (2013). https://doi.org/10.1038/nature12060

[6] Benedum, C. M., Shea, K. M., Jenkins, H. E., Kim, L. Y., and Markuzon, N. (2020). Weekly dengue forecasts in Iquitos, Peru; San Juan, Puerto Rico; and Singapore.

[7] Oladimeji Mudele, Alejandro C. Frery, Lucas F.R. Zanandrez, Alvaro E. Eiras, Paolo Gamba,'Modeling dengue vector population with earth observation data and a generalized linear model.'

[8] William Hoyos, Jose Aguilar, Mauricio Toro, Dengue models based on machine learning techniques: A systematic literature review.

[9] Choi Y, Tang CS, McIver L, Hashizume M, Chan V, Abeyasinghe RR, Iddings S, Huy R. Effects of weather factors on dengue fever incidence and implications for interventions in Cambodia. BMC Public Health. 2016 Mar 8;16:241. doi: 10.1186/s12889-016-2923-2. PMID: 26955944; PMCID: PMC4784273.

[10] Bakar, Azuraliza Abu et al. "2011 International Conference on Electrical Engineering and Informatics 17-19 July 2011 , Bandung , Indonesia Predictive Models for Dengue Outbreak Using Multiple Rulebase Classifiers." (2011).

[11] HWB Kavinga, DDM Jayasundara, and Dushantha NK Jayakody.A new dengue outbreak statistical model using the time series analysis. European International Journal of Science and Technology.

[12] Yusof, Yuhanis and Mustaffa, Zuriani. (2011). Dengue Outbreak Prediction: A Least Squares Support Vector Machines Approach. International Journal of Computer Theory and Engineering. 3. 489 493.10.7763/IJCTE.2011.V3.355.

[13] https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

[14] B. K.,. S. G. M. '. F. H. Sergio Ramírez-Gallego, A survey on data preprocessing for data stream mining: Current status and future directions, Elsevier, pp. 39-57, 2017.

[15] Patil S, Pandya S. Forecasting Dengue Hotspots Associated With Variation in Meteorological Parameters Using Regression and Time Series Models. Front Public Health. 2021 Nov 26;9:798034. doi: 10.3389/fpubh.2021.798034. PMID: 34900929; PMCID: PMC8661059.

[16] https://medium.com/@sreevalsan766/maximizing-model-performance-a-deep-dive-into-grid-search-cross-validation-22861d53a225