ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

Time Series Analysis-Based Prediction of Dengue Spread Using Climate Data

Sirimalle Srikanth[1], A. Shivani[2], B. Harshitha[2], B. Ganga Jyothi[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Electronics and Communication Engineering
[1,2]Malla Reddy Engineering College for Women, Maisammaguda, Hyderabad, Telangana, India

Abstract

Dengue is a human arbovirus disease transmitted by the female mosquito of the genus Aedes, mainly Aedes aegypti and Ae. albopictus. Dengue, the most frequent arthropod-borne viral disease, is prevalent in tropical and subtropical regions. Two major clinical forms of dengue illness involve the mild form of dengue fever and severe form mostly characterized by plasma leakage with or without haemorrhage. Dengue is endemic in all surrounding countries with the four serotypes circulating in the region within a period of ten years. Countries or territories with the highest number of reported dengue cases were Puerto Rico, the Dominican Republic, Martinique, Trinidad and Tobago and French Guiana. Population movement is an important factor in the virus dissemination. It contributes to carry new virus strains, but it also participates to introduce nonimmune subjects in an endemic area. This proposed system is built to predict the spread of dengue fever with climate data using the concept of time series analysis. In addition, this project also performs the exploratory data analytics on the dengue dataset over a period of time. Finally, prediction analysis also performed with the usage of advancement rendered by machine learning algorithms.

Keywords: Dengue fever, time series analysis, machine learning, predictive analysis.

1. Introduction

Dengue is a potentially life-threatening arboviral disease transmitted by female Aedes mosquitoes, especially A. aegypti, A. albopictus, and A. vitattus. These vectors are common tropical hematophagous ectoparasites. This zoonotic disease spread from African or Asian non-human primates 500 to 1000 years ago, but within the last 60 years it has spread from just 9 countries experiencing severe epidemics to become endemic in over 100 countries worldwide, even affecting non-tropical or subtropical areas. Moreover, approximately one hundred million people yearly suffer from the symptomatic disease caused by its four serotypes. Given the significant impact of environmental changes on disease transmission, the One Health approach is urgently needed to implement the integration between human, animal, and ecological health. The objective of this paper is to provide an insight into techniques that can be used for future predictive models based on the One Health perspective, particularly in respect to Latin America but also elsewhere.

One Health is a multidisciplinary approach that acknowledges the synergy between human and animal health and their shared environment. This idea is not new; the noted nineteenth-century pathologist (and originator of the term zoonosis) Rudolph Virchow famously asserted in 1858 that “between animal and human medicine, there are no dividing lines—nor should there be”. This approach has become increasingly important in the 21st Century with the convergence of the pressures of changing climate, migration of human and animal populations, and the growing human population that increases the proximity between wildlife and humans. Indeed, the term One Health was only coined in the early 2000s with the appearance of the zoonotic SARS and H5N1 influenza diseases.

Whilst the One Health perspective is widely seen as necessary and increasingly used for better disease control, epidemiological approaches have not kept up with this change. Conventional epidemiological perspectives tend to view disease broadly from a human-only perspective, focusing on human

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

demographic conditions with often only climatic/environmental factors accommodating the disease vector health. In contrast, One Health requires the health and lifecycle of the zoonotic disease vectors to be explicitly considered alongside the human environment, demographics, and interaction with the zoonotic host vectors.

For example, whilst environmental and sociological considerations often take a back seat in One Health, they frequently occupy the centre stage in epidemiology. Factors such as mean temperatures and rainfall used in predicting dengue, with a very vague consideration of how they affect the mosquito vectors, are an emergent challenge to be considered. High rainfall, for instance, is beneficial to mosquitoes because it provides water-filled locations for eggs and larvae, whilst the mosquitoes are primarily impervious to strikes by raindrops that might otherwise kill them. In addition, temperature and rain generally affect many other infectious and tropical diseases.

2. Literature Survey

Majeed, M.A.; Shafri, H.Z.M.; [1] dengue fever cases in Malaysia using machine learning techniques. A dataset consisting of weekly dengue cases at the state level in Malaysia from 2010 to 2016 was obtained from the Malaysia Open Data website and includes variables such as climate, geography, and demographics. Six different long short-term memory (LSTM) models were developed and compared for dengue prediction in Malaysia: LSTM, stacked LSTM (S-LSTM), LSTM with temporal attention (TA-LSTM), S-LSTM with temporal attention (STA-LSTM), LSTM with spatial attention (SA-LSTM), and S-LSTM with spatial attention (SSA-LSTM).

Cabrera, M.; Leake, J.; [2] epidemiological prediction of dengue fever using the One Health perspective, including an analysis of how Machine Learning techniques have been applied to it and focuses on the risk factors for dengue in Latin America to put the broader environmental considerations into a detailed understanding of the small-scale processes as they affect disease incidence. Determining that many factors can act as predictors for dengue outbreaks, a large-scale comparison of different predictors over larger geographic areas than those currently studied is lacking to determine which predictors are the most effective.

Dey, Samrat Kumar, et al. [3] develop a machine learning model that can use relevant information about the factors that cause Dengue outbreaks within a geographic region. To predict dengue cases in 11 different districts of Bangladesh, we created a DengueBD dataset and employed two machine learning algorithms, Multiple Linear Regression (MLR) and Support Vector Regression (SVR). This research also explores the correlation among environmental factors like temperature, rainfall, and humidity with the rise and decline trend of Dengue cases in different cities of Bangladesh. The entire dataset was divided into an 80:20 ratio, with 80 percent used for training and 20% used for testing. The research findings imply that, for both the MLR with 67% accuracy along with Mean Absolute Error (MAE) of 4.57 and SVR models with 75% accuracy along with Mean Absolute Error (MAE) of 4.95, the number of dengue cases reduces throughout the winter season in the country and increases mainly during the rainy season in the next ten months, from August 2021 to May 2022.

Kakarla, S.G., Kondeti, P.K., et al. [4] applied vector auto regression, generalized boosted models, support vector regression, and long short-term memory (LSTM) to predict the dengue prevalence in Kerala state of the Indian subcontinent. Consider the number of dengue cases as the target variable and weather variables viz., relative humidity, soil moisture, mean temperature, precipitation, and NINO3.4 as independent variables. Various analytical models have been applied on both datasets and predicted the dengue cases. Among all the models, the LSTM model was outperformed with superior prediction capability (RMSE: 0.345 and R2:0.86) than the other models.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

Roster, Kirstin, et al. [5] developed a model for predicting monthly dengue cases in Brazilian cities 1 month ahead, using data from 2007–2019. We compared different machine learning algorithms and feature selection methods using epidemiologic and meteorological variables. They found that different models worked best in different cities, and a random forests model trained on monthly dengue cases performed best overall. It produced lower errors than a seasonal naive baseline model, gradient boosting regression, a feed-forward neural network, or support vector regression.

Sarder, Faysal, et al. [6] predict the accuracy of dengue outbreak from climate data. A dengue dataset, containing information of climate variables, dengue cases during 2019 to 2021 from Meteorology Department and Directorate General of Health Services (DGHS), Bangladesh. We split the whole dataset into 70:30 ratios were 70% considered as training and 30% for testing purposes. Such, prediction of accuracy we apply various supervised machine learning (ML) algorithms like Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Naïve Bayes (NB), AdaBoostClassifier (AdaBoost), XGBRegressor, GradientBoostingClassifier and Random Forest (RF). Finally, from these algorithms, SVM provide the highest accuracy of 96.73%.

Ochida, N., Mangeas, M., et al. [7] proposed statistical estimation of the effective reproduction number (Rt) based on case counts to create a categorical target variable: epidemic week/non-epidemic week. A machine learning classifier has been trained using relevant climate indicators in order to estimate the probability for a week to be epidemic under current climate data and this probability was then estimated under climate change scenarios.

Anuranjan, M. B., et al. [8] considered three different modelling techniques: interpolation, gradient boosting regression and random forest regression. Parameters were tuned and adjusted for optimal performance. Results are based on prediction accuracy and mean absolute error (MAE). The performance was analysed, and the result points out that the gradient boosting regression performs significantly better than the other models and is therefore considered to be a better approach. Future results can be improved by obtaining large amounts of meaningful data and implementing better models associated with time series predicting.

Gupta, G.; Khan, S.; et al. [9] developed dengue predictive models, data from microarrays and RNA-Seq have been used significantly. Bayesian inferences and support vector machine algorithms are two examples of statistical methods that can mine opinions and analyze sentiment from text. In general, these methods are not very strong semantically, and they only work effectively when the text passage inputs are at the level of the page or the paragraph; they are poor miners of sentiment at the level of the sentence or the phrase.

3. PROPOSED SYSTEM

This research work combines time series analysis techniques, feature engineering, and machine learning using the XGBoost algorithm to predict dengue spread based on climate data. It contributes to early warning systems for dengue outbreaks and supports public health efforts to mitigate the impact of the disease. Figure 1 shows the proposed system model. The detailed operation illustrated as follows:

Step 1. Exploratory Data Analysis (EDA):

- **Data Collection**: Gather historical climate data and dengue spread records. This data typically includes variables such as temperature, humidity, rainfall, and the number of dengue cases over time.
- **Data Inspection**: Examine the dataset's structure, including its size, data types, and any missing values. Ensure that the data is correctly formatted for time series analysis.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

- **Time Series Decomposition**: Decompose the time series data into its constituent components, including trend, seasonality, and residual noise. This helps in understanding the underlying patterns in the data.
- **Data Visualization**: Create visualizations such as line plots, histograms, and seasonal decomposition plots to visualize the time series data. These visualizations can reveal trends, seasonality, and any anomalies.
- **Correlation Analysis**: Perform correlation analysis to identify relationships between climate variables (e.g., temperature, rainfall) and dengue spread. This helps in selecting relevant features for modeling.

Step 2. Preprocessing of Dataset:

- **Handling Missing Data**: Address any missing data points in the time series, using techniques like imputation or interpolation to fill gaps.
- **Feature Engineering**: Create additional features, such as lag variables (previous time steps), moving averages, and seasonality indicators, to capture relevant patterns in the data.
- **Normalization/Scaling**: Normalize or scale the data if necessary to ensure that all features have similar scales.
- **Train-Test Split**: Divide the time series data into training and testing sets. Typically, earlier data points are used for training, and more recent data points are reserved for testing.

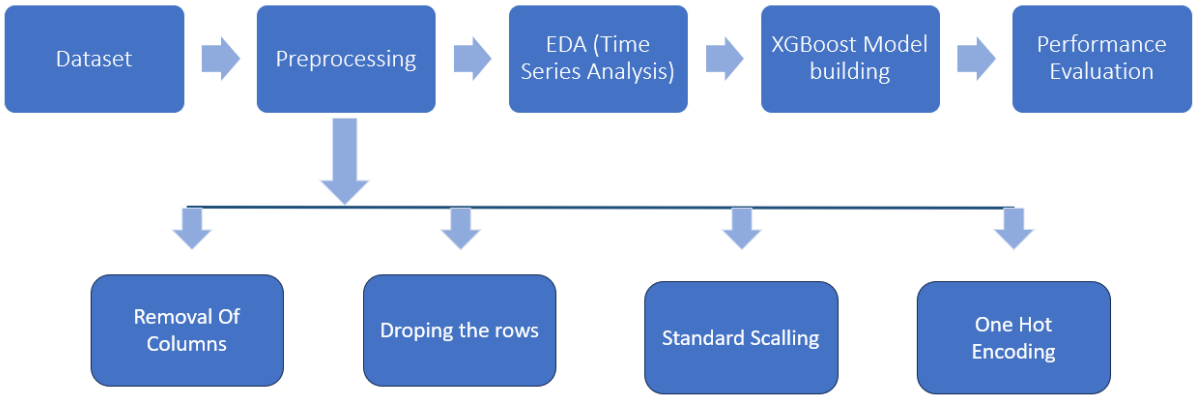

Figure 1. Proposed System model.

Step 3. XGBoost Model Training:

- **Selecting Features**: Choose the relevant climate variables and engineered features as input features (X) and the number of dengue cases as the target variable (y).
- **Hyperparameter Tuning**: Tune the hyperparameters of the XGBoost model, such as learning rate, maximum depth, and the number of estimators (trees), using techniques like grid search or random search.
- **Model Training**: Train the XGBoost model using the training data. XGBoost is a gradient boosting algorithm known for its effectiveness in time series forecasting.

Step 4. Prediction:

- **Model Evaluation**: Use the trained XGBoost model to make predictions on the test dataset. Evaluate the model's performance using appropriate metrics for time series forecasting, such as

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

- **Visualization of Predictions**: Visualize the model's predictions against the actual dengue cases on a time series plot. This allows you to assess how well the model captures the patterns and trends in the data.

4. RESULTS AND DISCUSSION

Figure 2 represents a portion of the dataset that was used for predicting the spread of Dengue fever. It includes various features (columns) and corresponding target values (total cases) for a specific period. Figure 3 is a line plot that displays the total number of Dengue fever cases over time for two different cities: San Juan and Iquitos. The x-axis represents time (likely in weeks or months), while the y-axis represents the total number of cases. There are two lines, one for each city, showing how the cases change over time.

	city	year	weekofyear	week_start_date	ndvi_ne	ndvi_nw	ndvi_se	ndvi_sw	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k
0	sj	1990	18	1990-04-30	0.122600	0.103725	0.198483	0.177617	12.42	297.572857	297.742857
1	sj	1990	19	1990-05-07	0.169900	0.142175	0.162357	0.155486	22.82	298.211429	298.442857
2	sj	1990	20	1990-05-14	0.032250	0.172967	0.157200	0.170843	34.54	298.781429	298.878571
3	sj	1990	21	1990-05-21	0.128633	0.245067	0.227557	0.235886	15.36	298.987143	299.228571
4	sj	1990	22	1990-05-28	0.196200	0.262200	0.251200	0.247340	7.52	299.518571	299.664286
...	...	...	...	...	...	...	...	...	...	...	...
1451	iq	2010	21	2010-05-28	0.342750	0.318900	0.256343	0.292514	55.30	299.334286	300.771429
1452	iq	2010	22	2010-06-04	0.160157	0.160371	0.136043	0.225657	86.47	298.330000	299.392857
1453	iq	2010	23	2010-06-11	0.247057	0.146057	0.250357	0.233714	58.94	296.598571	297.592857
1454	iq	2010	24	2010-06-18	0.333914	0.245771	0.278886	0.325486	59.67	296.345714	297.521429
1455	iq	2010	25	2010-06-25	0.298186	0.232971	0.274214	0.315757	63.22	298.097143	299.835714

1456 rows × 25 columns

reanalysis_tdtr_k	station_avg_temp_c	station_diur_temp_rng_c	station_max_temp_c	station_min_temp_c	station_precip_mm	total_cases
2.628571	25.442857	6.900000	29.4	20.0	16.0	4
2.371429	26.714286	6.371429	31.7	22.2	8.6	5
2.300000	26.714286	6.485714	32.2	22.8	41.4	4
2.428571	27.471429	6.771429	33.3	23.3	4.0	3
3.014286	28.942857	9.371429	35.0	23.9	5.8	6
...	...	...	...	...	...	...
9.800000	28.633333	11.933333	35.4	22.4	27.0	5
7.471429	27.433333	10.500000	34.7	21.7	36.6	8
7.500000	24.400000	6.900000	32.2	19.2	7.4	1
7.871429	25.433333	8.733333	31.2	21.0	16.0	1
11.014286	27.475000	9.900000	33.7	22.2	20.4	4

Figure 2: Sample dataset used for Dengue spread prediction.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

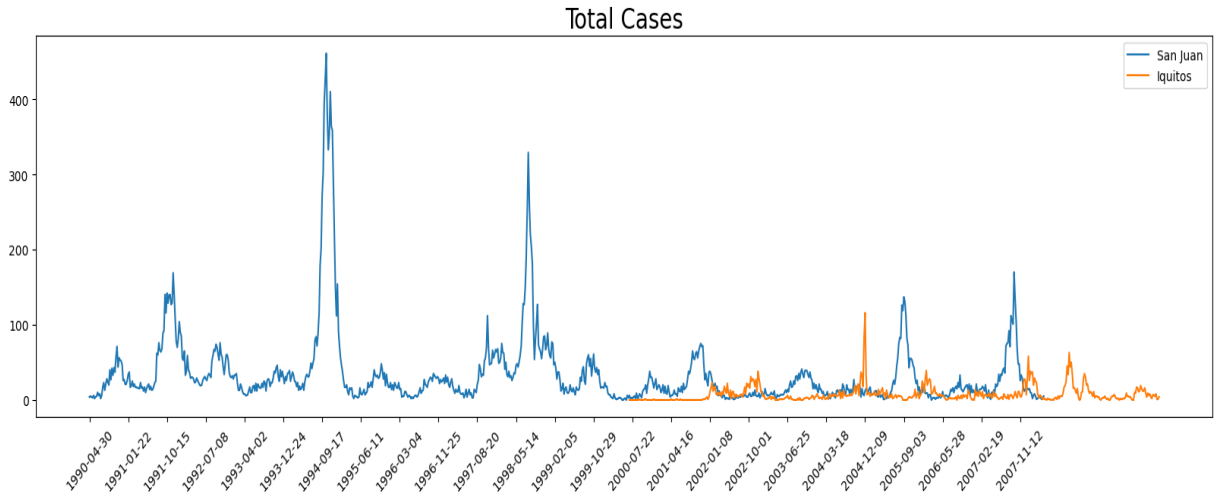

Figure 3: line plot to visualize the total number of Dengue fever cases over time for two different cities.

Figure 4 consists of multiple histograms, each representing the distribution of values for a numerical column in the DataFrame. It provides insights into the frequency of different values within each column.

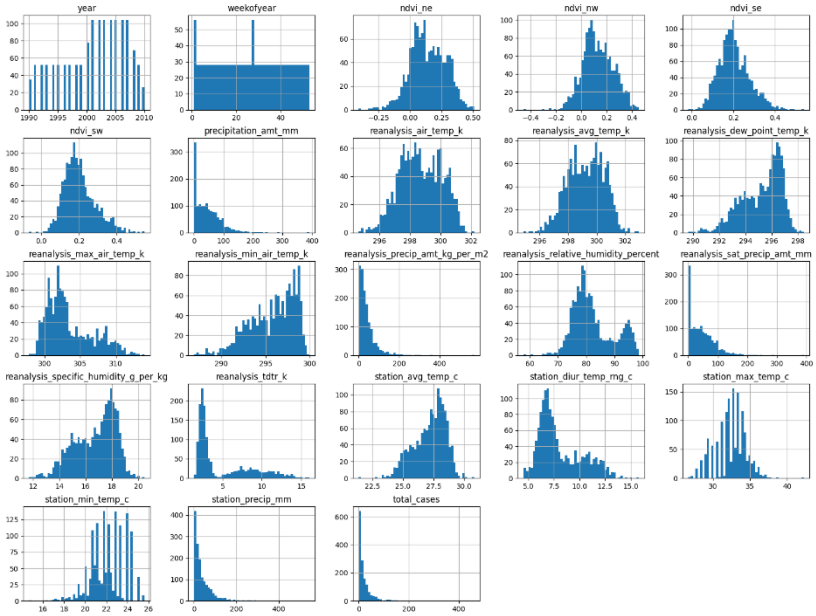

Figure 4: histogram for each numerical column in the Data Frame.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

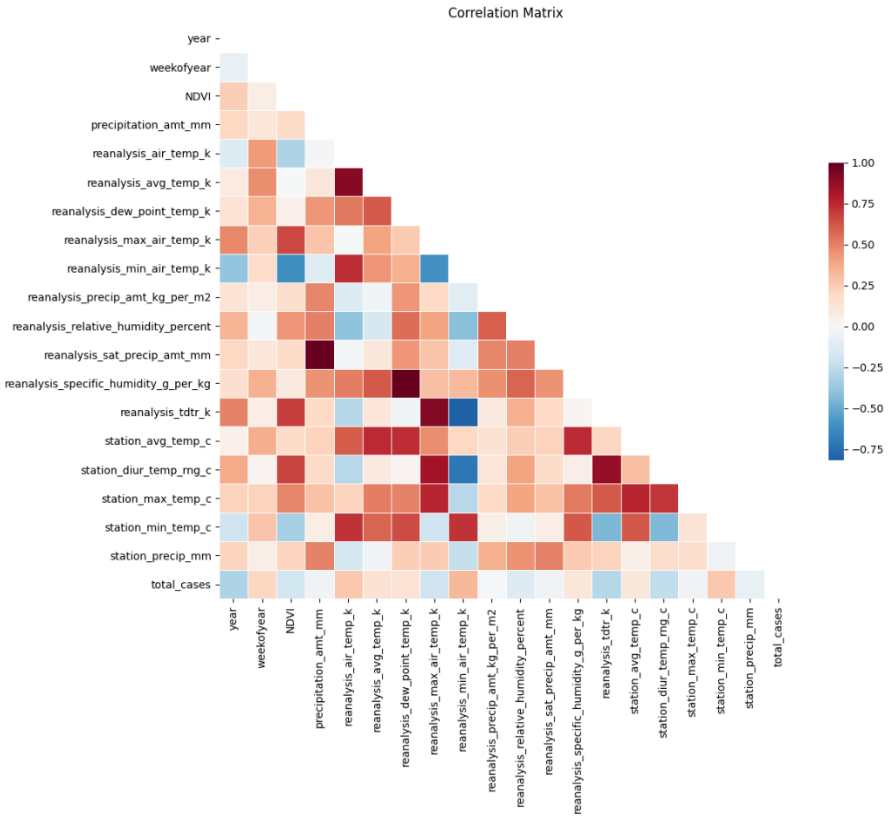

Figure 5: Heatmap of correlation of columns in a dataset used for dengue spread prediction.

Figure 5 is a heatmap that visualizes the correlation between different columns (features) in the dataset used for predicting Dengue spread. Each cell in the heatmap represents the correlation coefficient between two columns. A warmer color (closer to red) indicates a stronger positive correlation, while a cooler color (closer to blue) indicates a stronger negative correlation. Figure 6 displays a portion of the dataset after it has undergone preprocessing steps. Preprocessing may include tasks like handling missing values, feature engineering, and encoding categorical variables. It represents the cleaned and transformed data ready for modelling. Figure 7 specifically focuses on the features column(s) of the dataset after preprocessing. It may display the values, statistics, or distribution of the features that will be used for predicting Dengue spread.

	city	year	weekofyear	NDVI	week_start_date	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k	reanalysis_max_air_temp_k
0	1	1990.0	18.0	0.150606	1990-04-30	12.42	297.572857	297.742857	299.8
1	1	1990.0	19.0	0.157479	1990-05-07	22.82	298.211429	298.442857	300.9
2	1	1990.0	20.0	0.133315	1990-05-14	34.54	298.781429	298.878571	300.5
3	1	1990.0	21.0	0.209286	1990-05-21	15.36	298.987143	299.228571	301.4
4	1	1990.0	22.0	0.239235	1990-05-28	7.52	299.518571	299.664286	301.9
...	...	...	...	...	...	...	...	...	...
1451	0	2010.0	21.0	0.302627	2010-05-28	55.30	299.334286	300.771429	309.7
1452	0	2010.0	22.0	0.170557	2010-06-04	86.47	298.330000	299.392857	308.5
1453	0	2010.0	23.0	0.219296	2010-06-11	58.94	296.598571	297.592857	305.5
1454	0	2010.0	24.0	0.296014	2010-06-18	59.67	296.345714	297.521429	306.1
1455	0	2010.0	25.0	0.280282	2010-06-25	63.22	298.097143	299.835714	307.8

1456 rows × 21 columns

Figure 6: data frame after preprocessing used for dengue spread

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

	city	weekofyear	NDVI	precipitation_amt_mm	reanalysis_air_temp_k	reanalysis_avg_temp_k	reanalysis_max_air_temp_k	reanalysis_min_air_temp_k
936	0	26.0	0.228307	25.41	296.740000	298.450000	307.3	293.1
937	0	27.0	0.256012	60.61	296.634286	298.428571	306.6	291.1
938	0	28.0	0.170504	55.52	296.415714	297.392857	304.5	292.6
939	0	29.0	0.206918	5.60	295.357143	296.228571	303.6	288.6
940	0	30.0	0.316546	62.76	296.432857	297.635714	307.0	291.5
...	...	...	...	...	...	...	...	...
706	1	48.0	0.084369	16.20	298.814286	298.907143	301.0	297.2
707	1	49.0	0.073851	0.00	299.107143	299.242857	300.7	297.4
708	1	50.0	-0.007986	182.81	299.174286	299.185714	301.5	297.5
709	1	51.0	0.099544	0.00	298.555714	298.607143	300.5	296.9
710	1	52.0	0.071783	1.96	299.044286	299.178571	300.8	297.2

1049 rows × 18 columns

Figure 7: data frame of features column of a dataset after preprocessing

Figure 8 provides a view of the target column (total cases) of the dataset after preprocessing. It may show the distribution of target values and any transformations or adjustments made during preprocessing. Figure 9 is a line plot that compares the predicted cases (likely generated by a machine learning model) with the actual cases of Dengue fever for the city of Iquitos. The x-axis represents time, while the y-axis represents the number of cases. The plot helps assess how well the model's predictions align with the actual data. Figure 10 Similar to Figure 9, this figure compares the predicted cases with the actual cases of Dengue fever, but for the city of San Juan. Figure 11 displays the overall prediction results obtained using an XGBoost classifier. It may include metrics such as Mean Absolute Error (MAE) or other evaluation measures to assess the performance of the model.

	total_cases
1274	0.0
1275	14.0
1276	6.0
1277	10.0
1278	5.0
...	...
931	4.0
932	3.0
933	1.0
934	3.0
935	5.0

407 rows × 1 columns

Figure 8: Target column of a dataset after preprocessing

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

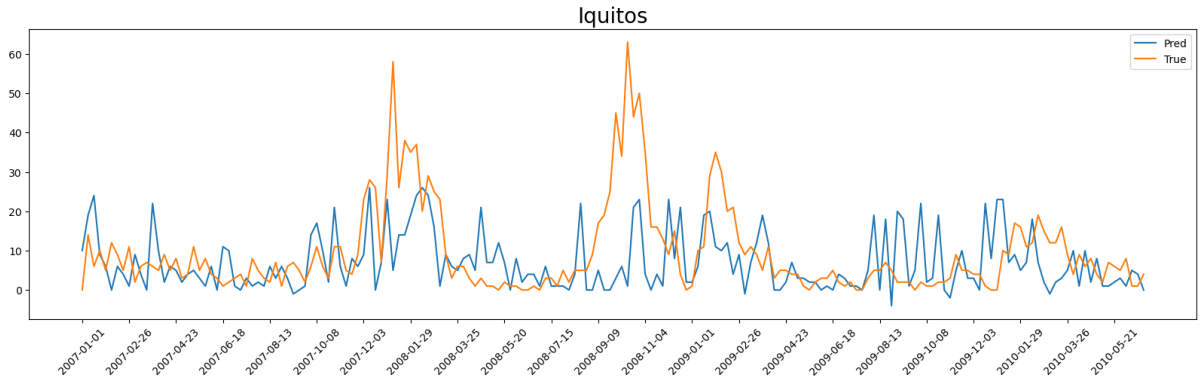

Figure 9: line plot to compare the predicted and actual cases of Dengue fever for the city of Iquitos (iq).

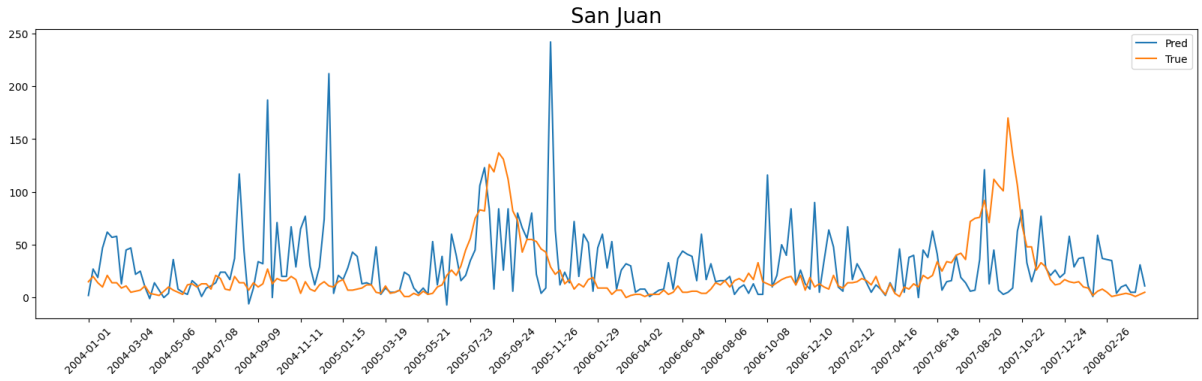

Figure 10: line plot to compare the predicted and actual cases of Dengue fever for the city of San Juan (S j).

	city	year	weekofyear	total_cases
0	sj	2008	18	3
1	sj	2008	19	34
2	sj	2008	20	23
3	sj	2008	21	36
4	sj	2008	22	16
5	sj	2008	23	21
6	sj	2008	24	4
7	sj	2008	25	38
8	sj	2008	26	3
9	sj	2008	27	4

Figure 11: Prediction results using XG Boost classifier.

5. CONCLUSION

In conclusion, this study embarked on the challenging task of predicting dengue fever, a highly volatile and rapidly evolving disease, by employing time series analysis with an ARIMA model. Traditional

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 12, Iss 4, 2023

dengue fever prediction methods often fall short due to its unique characteristics and lack of seasonality. However, through the application of machine learning techniques, we have explored patterns and trends within Dengue data, striving to gain insights into its future price movements. While the results may not always provide perfect predictions, they underscore the potential of machine learning models in understanding and forecasting dengue spread.

REFERENCES

[1] Majeed, M.A.; Shafri, H.Z.M.; Zulkafli, Z.; Wayayok, A. A Deep Learning Approach for Dengue Fever Prediction in Malaysia Using LSTM with Spatial Attention. Int. J. Environ. Res. Public Health 2023, 20, 4130. https://doi.org/10.3390/ijerph20054130

[2] Cabrera, M.; Leake, J.; Naranjo-Torres, J.; Valero, N.; Cabrera, J.C.; Rodríguez-Morales, A.J. Dengue Prediction in Latin America Using Machine Learning and the One Health Perspective: A Literature Review. Trop. Med. Infect. Dis. 2022, 7, 322. https://doi.org/10.3390/tropicalmed7100322

[3] Dey, Samrat Kumar, et al. "Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach." PLoS One 17.7 (2022): e0270933.

[4] Kakarla, S.G., Kondeti, P.K., Vavilala, H.P. et al. Weather integrated multiple machine learning models for prediction of dengue prevalence in India. Int J Biometeorol 67, 285–297 (2023). https://doi.org/10.1007/s00484-022-02405-z

[5] Roster, Kirstin, Colm Connaughton, and Francisco A. Rodrigues. "Machine-Learning–Based Forecasting of Dengue Fever in Brazilian Cities Using Epidemiologic and Meteorological Variables." American Journal of Epidemiology 191.10 (2022): 1803-1812.

[6] Sarder, Faysal, Sorefa Akter, and Sharmin Akter. "Predicting Dengue Outbreak from Climate Data Using Machine Learning Algorithms." 2022 IEEE International Conference on Data Science and Information System (ICDSIS). IEEE, 2022.

[7] Ochida, N., Mangeas, M., Dupont-Rouzeyrol, M. et al. Modeling present and future climate risk of dengue outbreak, a case study in New Caledonia. Environ Health 21, 20 (2022). https://doi.org/10.1186/s12940-022-00829-z.

[8] Anuranjan, M. B., et al. "Machine Learning Techniques for Predicting Dengue Outbreak." Innovations in Information and Communication Technologies: Proceedings of ICIICT 2022. Singapore: Springer Nature Singapore, 2022. 45-56.

[9] Gupta, G.; Khan, S.; Guleria, V.; Almjally, A.; Alabduallah, B.I.; Siddiqui, T.; Albahlal, B.M.; Alajlan, S.A.; AL-subaie, M. DDPM: A Dengue Disease Prediction and Diagnosis Model Using Sentiment Analysis and Machine Learning Algorithms. Diagnostics 2023, 13, 1093. https://doi.org/10.3390/diagnostics13061093.