Proceedings of the Fourth International Conference on Electronics and Sustainable Communication Systems (ICESC-2023)
IEEE Xplore Part Number: CFP23V66-ART; ISBN: 979-8-3503-0009-3

2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) | 979-8-3503-0009-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICESC57686.2023.10193045

Correlation based Feature Selection and Hybrid Machine Learning Approach for Forecasting Disease Outbreaks

Swayon Bhunia
UG Student, School of Computer Science & Engineering,
Vellore Institute of Technology- Chennai, India
swayon.bhunia2019@vitstudent.ac.in

Dr. Abirami S
Assistant Professor, School of Computer Science & Engineering,
Vellore Institute of Technology- Chennai, India
abirami.s@vit.ac.in

***Abstract*—According to WHO, Dengue is a viral infection transmitted to humans through the bite of infected mosquitoes i.e., Aedes aegypti mosquitoes. There is currently no known cure for dengue or severe dengue. Artificial Intelligence (AI) in the form of Machine Learning (ML) allows software programs to predict outcomes more correctly without explicit instructions. Machine learning algorithms use historical data as input to forecast new output values. The aim of this study is to identify, evaluate and interpret suitable hybrid algorithms/approaches relevant to the application of machine learning in limiting the spread of deadly disease outbreaks. It focuses on finding a way of predicting the next dengue fever local epidemic by comparing the bench mark approaches available until now. For this the study proposes the use of XGBoost coupled with Moving Average Rolling Features in order to learn the long-term temporal relations in the features to get accurate predictions. The dataset used for evaluating the proposed approach contains number of cases in the two locations: San Juan and Iquitos and it includes information on temperature, precipitation, humidity, vegetation, and what time of the year the data was obtained. A correlation analysis-based feature selection along with Moving Average Rolling Features has been used for getting more precise data implemented with ML approach resulting in MSE 11.37 in San Juan and MSE 6.37 in Iquitos.**

Keywords— Dengue, XGBoost, Machine Learning (ML), Moving Average Rolling.

I. INTRODUCTION

Dengue viruses spread to people when an infected Aedes species mosquito bites a person. Nearly half of the world's population, or 4 billion people, live in dengue-risk areas. Dengue is frequently the primary cause of illness in high-risk areas. A viral disease called dengue virus disease, also known as dengue fever or simply "dengue," is carried by mosquitoes and transmitted in many tropical as well as subtropical regions of the world, including Africa, Asia, South America, and sporadically some areas of northern Queensland.

DF is a dynamic, systemic infection that manifests a variety of clinical signs, both serious and not so serious. The flu-like symptoms of fever, rash, and muscle and joint discomfort are present in mild instances. According to the WHO (2009), severe dengue fever episodes can result in severe bleeding, low blood pressure, and even fatality.

Currently, there is no cure for dengue; the only treatment is symptomatic and involves intensive patient care. Therefore, making predictions about impending epidemics can be quite helpful.

Research in Machine Learning (ML) focuses on comprehending and developing "learning" strategies, or techniques that use data to enhance performance on a certain set of tasks. It is considered to be a component of artificial intelligence. Without being explicitly instructed to do so, machine learning algorithms build a model from sample data, commonly referred to as training data, in order to make predictions or judgements. Machine learning algorithms are employed in a range of industries when it is challenging or impractical to create conventional algorithms that can perform the required tasks, including computer vision, speech recognition, etc. Among some studies, one of which used ML and deep learning algorithms along with a lexicon-based method to achieve accuracy.

It is possible to use machine learning (ML) techniques to stop the spread of severe infectious disease outbreaks, like dengue. This can be achieved by using machine learning techniques for the detection and prediction of lethal infectious illnesses.

Since mosquitoes are the primary vectors of dengue fever, climate factors, particularly temperature and precipitation/humidity, have a strong correlation with dengue fever transmission. The goal of the study is to forecast the number of dengue cases each week (in San Juan and Iquitos, respectively) using all the environmental factors (temperature, precipitation, vegetation, and more) that have been provided.

Most evaluations mislaid from their discussions the machine learning techniques, datasets, and performance metrics that were employed in a variety of applications for predicting and diagnosing the deadly infectious disease. So, by utilizing a variety of new methods, accuracy is increased while reducing processing time for massive datasets, which will, in my opinion, improve the current system.

The algorithms now in use for the system are outdated versions with poor accuracy and lengthy processing times. For improved outcomes, more sophisticated execution is required . As a result, proposed research work completely depends on excellent accuracy and quick processing.

II. RELATED WORKS

Machine learning is becoming increasingly popular because to modern technology, and it is growing swiftly. Without realizing it, we utilize machine learning on a daily basis in programmes like Alexa, Google Assistant, and Maps. Image recognition, speech recognition, traffic forecasting, product recommendations, self-driving cars, are some of the most popular real-world applications of machine learning.

In paper [1], for the purpose of predicting dengue, statistical models (VAR) and machine learning and deep learning models were used by author Satya Ganesh Kakarla. The LSTM model predicted monthly dengue cases using weather variables and dengue cases, outperforming all other models.

In Paper [2], authors Rajasekhar Mopuri, employed multi-step polynomial regression models and seasonal autoregressive integrated moving average (SARIMA) to forecast the number of malaria cases in the study area. The polynomial model showed that the population and malaria cases at lag one played a significant effect in malaria transmission and had a good prediction value for malaria cases.

A climate-driven dengue model was created by the authors of Paper [3] and predicted locations in India that would be susceptible to dengue transmission under current and future climate change scenarios. The study's author, Satya Ganesh Kakarla, also predicted the dengue dispersion risk map using representative concentration paths for India for the years 2018-2030 (near future), 2031-2050 (mid-future), and 2051-2080 (far future).

An ML-based model to predict dengue disease was proposed by Dhiman Sharma, in Paper [4]. Hospitals affiliated with the Dhaka and Chittagong medical colleges provided information on 209 patients. 23 highlighted datasets were created using preprocessed patient personal information, clinical information, and diagnostic information. The dataset was subjected to the random forest (RF) and decision tree (DT) algorithms. Finally, the decision tree's accuracy of 79% allowed us to finalise the method for classifying three different forms of dengue fevers in our model.

According on socioeconomic factors and machine learning (ML) algorithms, writer Phani Krishna Kondeti, estimated the occurrence of filariasis in Paper [5].

Using GR feature selection and 400% oversampling, NB produced the best AUC (64%) of any model. In a similar vein, J48 produced 23 and had an AUC of 62%. The impact of different causal factors on dengue fever outbreaks in the Delhi metropolitan area has been examined in Paper [6]. Lower error statistics (MAE, and standard errors) and a higher r squared correlation value indicate the presence of significant causal factors. Results by Shuchi Mala and Mahesh Kumar Jat show a significant correlation between DF occurrences and temperature, humidity, wind speed, daylight hours, age, built-up density, vegetation density, and distance from dairy locations, waterbodies, and drainage networks..

Paper [7] cites Salvador Gomez-Carro, completed the work. Using a nonlinear neural network approach to model dengue fever, it is possible to forecast dengue fever incidence rates in Mexico and Puerto Rico with a power of more than 70%. Precipitation, population size, air temperature, prior dengue cases, and date were found to be the most significant predictors of dengue fever epidemic occurrences in four model runs, two for population at risk and two for most vulnerable population in Yucatan, Mexico and San Juan, Puerto Rico.

The potential of Random Forest and its excellent predictive capabilities are shown in Paper [8] by Janet Ong, Xu Liu, in stratifying the spatial risk of dengue transmission in Singapore. An accurate dengue risk map created using Random Forest can help direct vector control efforts and enable for targeted preventive actions both before and during dengue outbreaks.

The LSTM ensemble forecasting method is shown in Paper [9] to be capable of accurately capturing the dynamic behavior of real-world time series. The proposed method by authors Jae Young Choi and Bumshik Lee outperforms other well-known forecasting algorithms, according to comparative studies on the four difficult time series. Authors Shalini Gambhir, Sanjay Kumar Malik, and Yugal Kumar state in Paper [10] that the ANN technique outperforms DT and NB while requiring a longer computing time.

Thus, it can be said that of the three machine learning methods, deep learning methods including XGBoost, LSTM, and ANN-based diagnostic model are all suitable for accurately detecting dengue sickness at an early stage. Finally, the papers use several algorithms and their combinations, primarily the older ones, to obtain increasingly accurate and precise predictions.

There is currently no effective treatment for dengue; instead, patients must receive extensive patient care. This is why making predictions about impending epidemics might be helpful. Since mosquitoes are the primary vectors of DF, climate factors, particularly temperature and precipitation/humidity, have a strong correlation with DF transmission.

The task is to predict the number of dengue cases each week in any of the two locations based on all the environmental variables provided (temperature, precipitation, vegetation, and more).

The algorithm already used for the current system is an old version algorithm, with low accuracy and more processing time. So, we need more sophisticated execution for better results by carrying out comparison between best machine learning approaches and deep learning.

III. Proposed Methodology

A. Data Collection

- Direct downloads of the project's data were made at DrivenData.org. The data includes details on temperature, precipitation, humidity, vegetation, and the time of year the data was collected in addition to the total number of instances in the two sites as shown in Table 1 and Table 2.
- dengue_features_train.csv (1,457 rows) containing data for both San Juan (years ranging from 1990 to 2008) and Iquitos (years ranging from 2000 to 2010), plus all the relevant features, for a total of 24 columns.
- dengue_features_test.csv (417 rows) containing data for both San Juan (years ranging from 2008 to 2013) and Iquitos (years ranging from 2010 to 2013), plus all the relevant features, for a total of 24 columns.

TABLE 1 SAMPLE VALUES FOR EVERY FEATURE IN THE DATASET

Features	Sample Values
city	sj
year	1990
weekofyear	18
week_start_date	30/04/1990
Normalize difference Vegetation Index_ne	0.1226
Normalize difference Vegetation Index_nw	0.103725
Normalize difference Vegetation Index_se	0.1984833
Normalize difference Vegetation Index_sw	0.1776167
preci_amt_mm	12.42
reanalysis_air_temp_k	297.572857143
reanalysis_avg_temp_k	297.742857143
reanalysis_dew_point_temp_k	292.414285714
reanalysis_max_air_temp_k	299.8
reanalysis_min_air_temp_k	295.9
reanalysis_precip_amt_kg_per_m2	32.0
reanalysis_relative_humidity_percent	733.657142857
reanalysis_sat_precip_amt_mm	12.42
reanalysis_specific_humidity_g_per_kg	140.128571429
reanalysis_tdtr_k	262.857142857
st_avg_temp_c	254.428571429
st_diur_temp_rng_c	**6**.9
st_max_temp_c	29.4
st_min_temp_c	20.0
st_precip_mm	16.0

B. Data Preprocessing

a) Missing Values: Missing values are given for the datasets from both cities during the exploratory data analysis that was done on the data. Since missing values are a typical occurrence in data, despite the fact that most predictive modelling algorithms cannot manage them, all missing values at this stage are imputed with the mean of the features in the datasets itself. To prevent any potential data leaking issues, mean values are only calculated on the train set and those values are used to perform imputation for the test set as well.

b) Outliers: In addition to missing data, outliers are also examined during the exploratory analysis and dealt with using the flooring and capping procedures based on the tenth and ninetieth percentiles of the distribution of the characteristics in the trainsets of the two cities, respectively. The test sets are also floored and capped using those values.

c) Features Selection: High correlation features are more linearly dependant and hence virtually equally affect the dependent variable. We can thus exclude one of the two features when there is a substantial correlation between two features. On the basis of exploratory data analysis, features are chosen. Only one of the first features is retained when there is only one left after removing those with strong pairwise correlation (0.90). Once more, the train sets for the two cities are used for this analysis, which is then transmitted to the test sets. The train and test sets also no longer contain the columns Normalize difference Vegetation Index_sw and Normalize difference Vegetation Index_se, which for the San Juan data turned out not to be connected to the target variable

TABLE 2 FEATURES SELECTED ARE INDICATED WITH 'X'

Features	San Juan	Iquitos
city	X	X
year	X	X
weekofyear	X	X
week_start_date	X	X
Normalize difference Vegetation Index_ne	X	X
Normalize difference Vegetation Index_nw	X	X
Normalize difference Vegetation Index_se		X
Normalize difference Vegetation Index_sw		X
precipitation_amt_mm		
reanalysis_air_temp_k		X
reanalysis_avg_temp_k		
reanalysis_dew_point_temp_k		
reanalysis_max_air_temp_k		
reanalysis_min_air_temp_k	X	
reanalysis_precip_amt_kg_per_m2		
reanalysis_relative_humidity_percent		
reanalysis_sat_precip_amt_mm	X	X
reanalysis_specific_humidity_g_per_kg	X	X
reanalysis_tdtr_k	X	X

i. Exploratory Data Analysis

One thing to consider in such a dataset is whether the two cities under consideration, San Juan (Puerto Rico) and Iquitos (Peru), contain significant differences in their distribution of variables representing vegetation, temperature, and humidity. This is after performing a first exploratory analysis on basic statistics for both the train and test set. Figure 1 shows the four charts that illustrate how four of the train set's most important variables behave differently depending on the city.

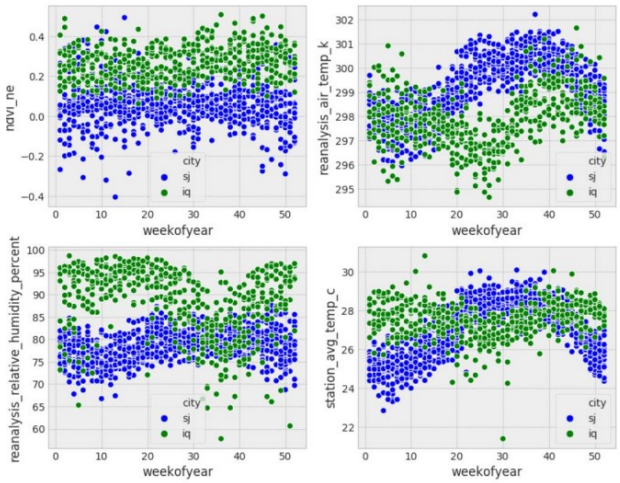


Figure 1: selected only representative variables

The distributions for the two cities appear to be significantly dissimilar from one another, as one can observe. Most likely because Iquitos is in the Southern hemisphere and San Juan is in the Northern hemisphere. In fact, all of the factors affecting vegetation distribution as well as those affecting humidity, temperature, and precipitation appear to behave in almost opposite ways in the two cities depending on the week of the year. As a result, the two cities will now be handled differently and fitted with various models.

ii. Feature Correlation

The train sets for San Juan and Iquitos are significantly connected in the two plots below. To prevent any issues with data leaking and to reduce model complexity, these features will be eliminated in the Data Pre-processing Section before being replicated in the test set. The decision to keep the two cities apart is supported by the differences in the correlations between these two cities shown in Figure 2.

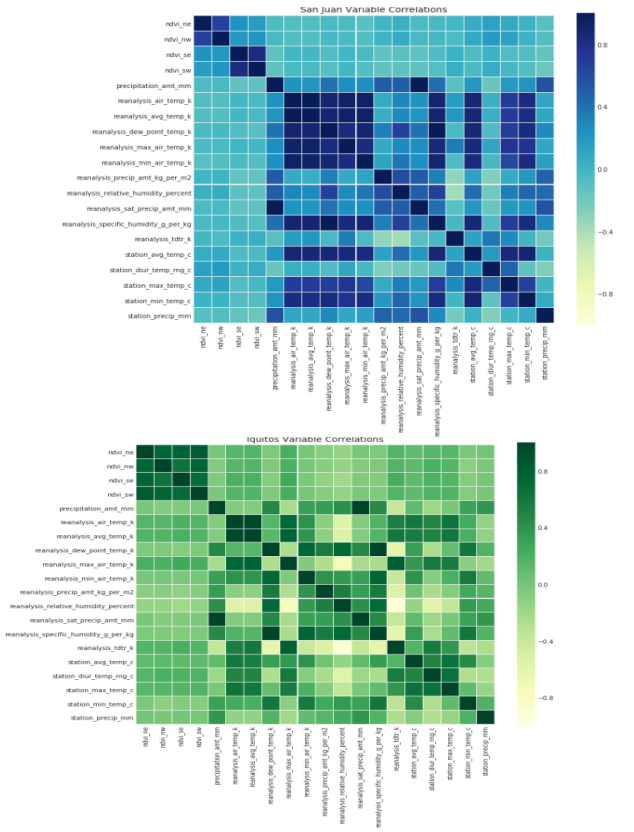


Figure 2: highly-correlated variables

iii. Target variable distribution and correlations

The historical history of the total number of cases in the two cities across the various years considered is

combined in the following graphs shown in Figure 3, along with their corresponding relative frequencies.

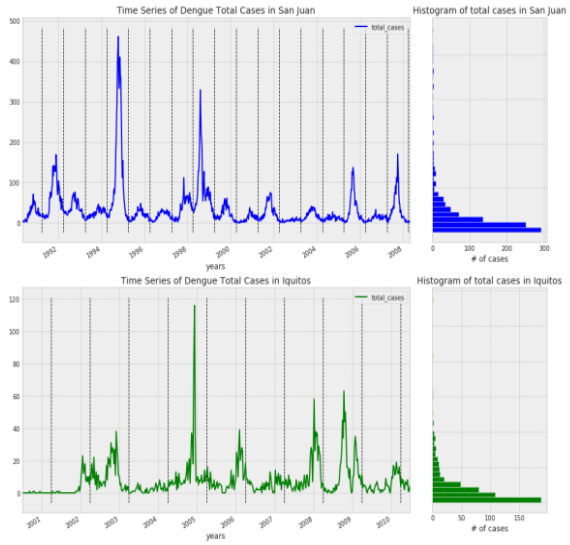
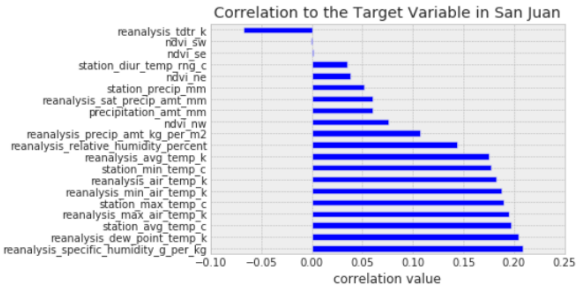
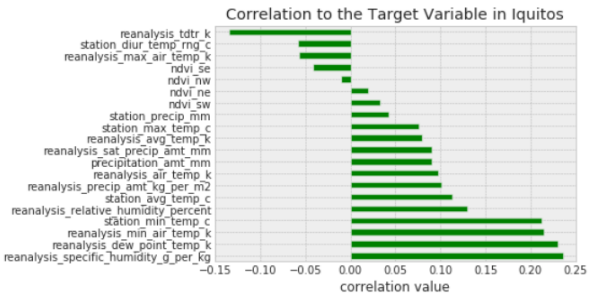


Figure 3: time series of dengue total cases

As we can see, San Juan experienced many peaks in cases in 1995, 1999, followed by two more mild increases in 2006 and 2008, respectively. In contrast, Iquitos experienced its biggest peak of infections in 2005, a year in which San Juan experienced very few cases, and was subsequently followed by 2009 and 2008.



(a)



(b)

Figure 4(a) Correlation to the target variable in San juan (b) Correlation to the target variable in Iquitos

Figure 4(a) and (b) shows the correlation to the target variable in San juan and Iquitos respectively. All of the aforementioned data will be examined and used to manage outliers in the train and test set. When deciding how to cap or floor the data in the train and test sets by removing certain rows, it will be looking for outliers in the train set.

iv. Model building

Extreme Gradient Boosting (XGBoost) with Moving Average Rolling Features: In addition to handling missing values properly, performing well on small datasets, and avoiding overfitting, XGBoost leverages parallel processing for quick speed. With all of these benefits, XGBoost is a well-liked remedy for regression issues like predicting. Although XGBoost is primarily intended for classification and regression tasks, time series forecasting issues can also be solved with it because of its scalability in all conditions.

XGBoost is the first model to be put into practice. Eighty percent of the train data for each city is preserved as the training set, while the remaining ten percent is used as the validation set. To fit the model to the data, the xgboost package's XGBRegressor function is utilized.

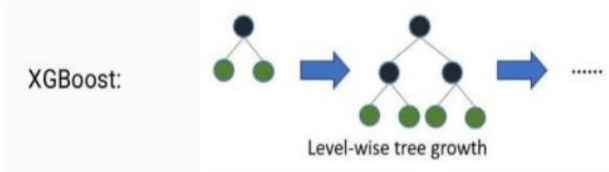


Figure 5: xgboost working mechanism. Image is taken from: www.anyscale.com

The goal of boosting is to improve each subsequent model by addressing the shortcomings of the one that came before it. The following models are built on top of the prior model as shown in Figure 5. The individual models might not be effective for the entire dataset, but they are effective for some portions of it. Because each model improves upon the performance of the one before it, the name is fitting.

It looks like the predictions are a little bit "late" in detecting peaks in total cases. Here we will try to add Moving Average Rolling Features to check for improvements which is the novelty of the approach.

v. Algorithms and techniques

Given that we are aware that the goal variable is a discrete numeric variable (the number of cases), the task of forecasting the total number of dengue cases for the upcoming X weeks in the future is a supervised issue (more particularly, a time series problem). Because of this, any forecasting technique or model that can accept features of many types as an input may be suitable.

Using rolling and moving averages to analyse data for a particular time series and identify trends. Use a longer

period of time when viewing these averages on a line chart to highlight long-term patterns.

Rolling Average: A rolling average updates the average of a data set continually to take into account all of the data up to that point. For instance, summing the return amounts from all 7 weeks and dividing the total by 7 would yield the rolling average of return quantities at the 7th week.

The average of a set of data is computed for a given period using a moving average. For instance, summing the return quantities over a given period of two weeks and dividing the result by two would yield the moving average of return quantities at the seventh week with a set period of two.

vi. Architecture

The proposed methodology and its workflow of how the project works and how the collected data is being used to predict using ML are interpreted in Figure 6.

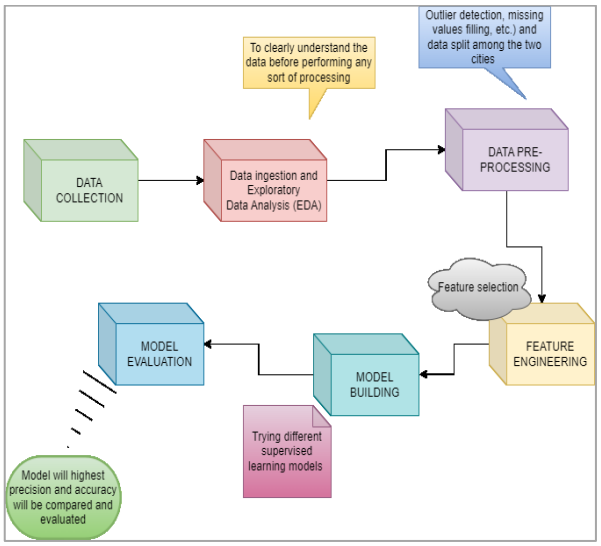


Figure 6: Methodology diagram

- Data Collection & Exploratory Data Analysis **:** Direct downloads of the project's data were made at DrivenData.org. The data includes details on temperature, precipitation, humidity, vegetation, and the time of year the data was collected in addition to the total number of instances in the two sites. Do the two cities differ from one another? Target Variable Distribution and Correlations, Features Correlations.

- Feature Engineering:
  Missing Values, Outliers: Features Selection, calculating mean, average of certain features.

- Model Building & Evaluation:
  Hybrid implementations of the Moving and Rolling Average Feature in XGBoost. Mean Absolute Error (MAE), a model evaluation statistic frequently employed with regression models, is one of the evaluation metrics.

Results and Discussion

Starting with data preparation, look for missing values and substitute the features' own means for them. In order to prevent any potential data leaking, the mean values were applied to both the train set and the test set. Both cities receive this treatment.
The actual vs. projected values plot is one of the other graphic evaluation factors that is taken into account.
Without any additional features, the first model, XGBoost, is built by merely fine-tuning the hyperparameters. The real vs. anticipated plot for the city of San Juan may be found in Figure 7, and the resulting MAE is 17.84. It appears that the projections are a bit slow to identify peaks in the number of occurrences overall.

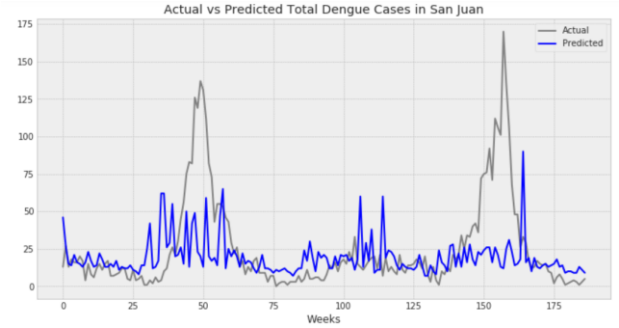


Figure 7: actual vs predicted total dengue cases in san juan using xgboost (no features included)

Moving average rolling elements are included in the model to check for advancements. Rolling means at 7 and 14 weeks are added to each feature that has been selected for the city after doing various testing. Figure 8's real vs. anticipated plot shows improved results with an MAE of 11.49.

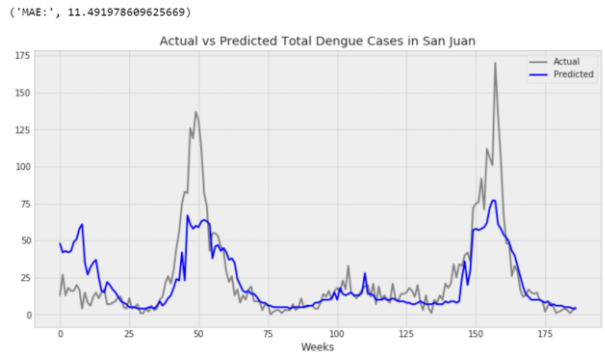


Figure 8: actual vs predicted total dengue cases in san juan using xgboost (rolling avg. Feature included)

Regarding Iquitos, The resulting MAE is 7.06, and Figure 9 shows the actual vs. expected plot. In this instance, the model is obviously too conservative because it fails to detect the peaks in the number of dengue cases overall while also appearing to be overfitting. In order to assess any advancement, moving average rolling characteristics are added to the model. Each feature selected for the city is given rolling means at 1, 5, 10, and 15 weeks after doing some tests, including on the hyperparameters.

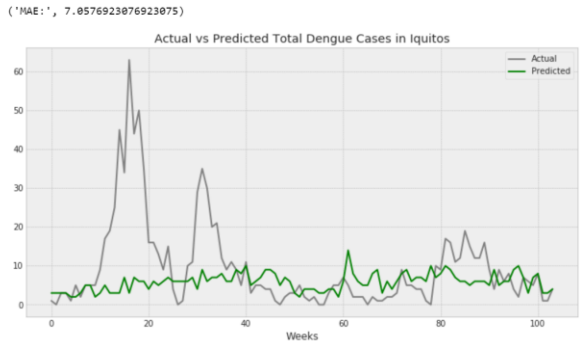


Figure 9: actual vs predicted total dengue cases in iquitos using xgboost (no features included)

Results improve though not much, showing a MAE of 6.37, actual vs predicted plot can be found in Figure 10.

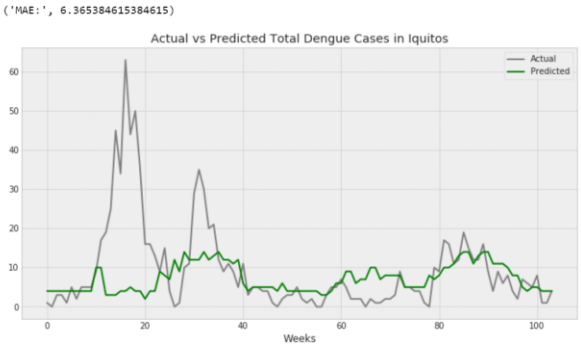


Figure 10: actual vs predicted total dengue cases in iquitos using xgboost (features included)

The final step is to go through every step again since I want to adhere strictly to the benchmark model. Making a function that gathers data and performs pre-processing on it comes first.

The actual vs. anticipated plot for San Juan City, where the MAE is 20,64, is shown in Figure 11. In reality, the peaks in the city of San Juan escape the model. The predictions in the actual vs expected plot show that the total predicted dengue cases have a fairly cyclical pattern.

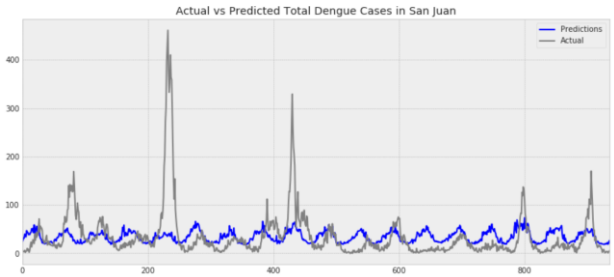


Figure 11: actual vs predicted total dengue cases in san juan using benchmark model

Figure 12 displays the actual vs. projected plot for the city of Iquitos, where the resulting MAE is 6.53. The model appears to detect trends more accurately than the benchmark model fitted for the city of San Juan, but it is still far from being capable of accurately forecasting peaks.

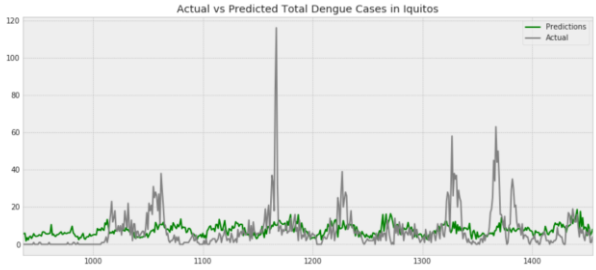


Figure 12: actual vs predicted total dengue cases in iquitos using benchmark model

Results for models fit to data from San Juan and Iquitos cities are summarized in the Table 3. As for the XGBoost model, only the better of the two - i.e., the one containing the moving average rolling features - is considered.

TABLE 3: COMPARISON BETWEEN ALL MODELS

City	Model	MSE (train set)
San Juan	XGBoost	17.84
	XGBoost (feature included)	11.49
	Benchmark Model (Negative Binomial)	20.64
Iquitos	XGBoost	7.06
	XGBoost (feature included)	6.37
	Benchmark model (Negative Binomial)	6.53

When only the MSE is taken into account, the XGBoost with Moving Rolling Feature for both cities seems to be the most advantageous model. The most noticeable difference, however, may be found in San Juan, where the feature-rich version of XGBoost performs considerably better at projecting overall future dengue cases. The models appear to perform relatively similarly when only

the MAE is considered in the city of Iquitos, where the difference is minimal.

Because of this, it is essential to analyse actual vs. anticipated graphs for each of the many models. Comparing the charts to the other models, it is much simpler to observe how effectively XGBoost predicted the phenomenon in both places.

Benchmark Model (Negative Binomial): Compared to a standard linear regression model, these models have a number of advantages, such as a skew, discrete distribution, and a constraint on the range of predicted values to non-negative numbers. A variation of the Poisson distribution known as the negative binomial distribution treats the distribution's parameter as a random variable in and of itself. A variance of the data that is greater than the mean can be explained by the variation of this parameter.

Regarding the benchmark model, the model has some flaws: the timing of the seasonality of the forecasts is out of sync with the actual findings, and the predictions are relatively (too) consistent, missing the spikes that signify the major outbreaks (thus the most dangerous times).

IV. CONCLUSION

The total number of dengue cases in San Juan (Puerto Rico) and Iquitos (Per) have been predicted using a variety of statistical and HYBRID machine learning methods in this study, and it is been compared to a benchmark model. Data intake and Exploratory Data Analysis (EDA) were carried out during the analysis in order to fully comprehend the data prior to any sort of processing. Following that, data were pre-processed and ready for the models. XGBoost and Negative Binomial Regression, used as a benchmark model, are the models taken into account. The XGBoost with moving average rolling features has been shown to be the method with the best performances (MAE and real versus expected values).

Hence, with respect to the benchmark model, seasonality timing mismatch issue has been addressed, along with the issue of not predicting spikes at all.

REFERENCES

[1] Satya Ganesh Kakarla, Phani Krishna Kondeti, Hari Prasad Vavilala , Gopi Sumanth Bhaskar Boddeda, Rajasekhar Mopuri, Sriram Kumaraswamy, Madhusudhan Rao Kadiri, Srinivasa Rao Mutheneni (2022). Weather integrated multiple machine learning models for prediction of dengue prevalence in India.

[2] Rajasekhar Mopuri, Satya Ganesh Kakarla, Srinivasa Rao Mutheneni, Madhusudhan Rao Kadiri, Sriram Kumaraswamy (2020). Climate based malaria forecasting system for Andhra Pradesh, India.

[3] Satya Ganesh Kakarla a, Kantha Rao Bhimala b, Madhusudhan Rao Kadiri a, Sriram Kumaraswamy a, Srinivasa Rao Mutheneni (2020). Dengue situation in India: Suitability and transmission potential model for present and projected climate change scenarios.

[4] Dhiman Sharma, Sohrab Hossain, Tanni Mittra, Md. Abdul Motaleb Bhuiya, Ishita Saha, Ravina Chakma (2020). Dengue Prediction using Machine Learning Algorithms

[5] Phani Krishna Kondeti, Kumar Ravi, Srinivasa Rao Mutheneni, Madhusudhan Rao Kadiri, Sriram Kumaraswamy, Ravi Vadlamani, Suryanaryana Murty Upadhyayula (2019). Applications of machine learning techniques to predict filariasis using socio-economic factors.

[6] Shuchi Mala, Mahesh Kumar Jat (2018). Implications of meteorological and physiographical parameters on dengue fever occurrences in Delhi.

[7] Abdiel E. Laureano-Rosario, Andrew P. Duncan, Pablo A. Mendez-Lazaro, Julian E. Garcia-Rejon, Salvador Gomez-Carro, Jose Farfan-Ale, Dragan A. Savic, Frank E. Muller-Karger (2018). Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico.

[8] Janet Ong, Xu Liu, Jayanthi Rajarethinam, Suet Yheng Kok, Shaohong Liang, Choon Siang Tang, Alex R. Cook, Lee Ching Ng, Grace Yap (2018), Mapping dengue risk in Singapore using Random Forest.

[9] Jae Young Choi, Bumshik Lee (2018), Combining LSTM Network Ensemble via Adaptive Weighting for Improved Time Series Forecasting.

[10] Shalini Gambhir, Sanjay Kumar Malik, Yugal Kumar (2018), The Diagnosis of Dengue Disease: An Evaluation of Three Machine Learning Approaches.

[11] Brownlee, J., 2020. A Gentle Introduction to Long Short-Term Memory Networks by the Experts, s.l.: Machine Learning Mastery.

[12] Brownlee, J., 2020. How to Use XGBoost for Time Series Forecasting, s.l: Machine Learning Mastery. C., S. & G.I., W., 2011. Encyclopedia of Machine Learning., Boston, MA: s.n.

[13] DrivenData, n.d. DrivenData. [Online] Available at: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

[14] Guestrin, C. & Chen, T., 2016. XGBoost: A Scalable Tree Boosting System, s.l.: ArXiv.Org. Hochreiter, S. & Schmidhuber, J., 1997. Long Short-Term Memory, s.l.: Neural Computation.

[15] Kuhn, M. & Johnson, K., 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models (Chapman & Hall/CRC Data Science Series). 1 ed. s.l.:Chapman and Hall/CRC..

[16] M., P., 2020. Illustrated Guide to LSTM's and GRU's: A step by step explanation, s.l.: Medium.

[17] [Organization, W. H., 2009. Dengue: guidelines for diagnosis, treatment, prevention and control. [Online] Available at: https://www.who.int/tdr/publications/documents/dengue-diagnosis.pdf

[18] Reimers, N. & Gurevych, I., 2017. Optimal hyperparameters for deep lstm-networks for sequence labeling tasks, s.l.: s.n.

[19] T.S., S., T., D. E. S.-G. & E.S.L., D. A., 2018. History, epidemiology and diagnostics of dengue in the American and Brazilian contexts: a review.. Parasites Vectors, Issue 11, p. 264.

[20] Wikipedia, 2022. Wikipedia. [Online] Available at: https://it.wikipedia.org/wiki [Accessed 03 2021].