Forecasting Dengue Cases by using Machine Learning Considering Meteorological Feature

Joy Barai[1], Maleha Israt Chowdhury[2],
Kamarum Monira Mow[3] and Shamim Ripon[4]

East West University, A/2 Jahurul Islam Ave, Dhaka 1212, Bangladesh
[1] 2018-1-60-011@std.ewubd.edu, [2] 2018-1-60-015@std.ewubd.edu
[3] 2018-1-60-016@std.ewubd.edu, [4] dshr@ewubd.edu

Abstract. Dengue is one of the most dangerous fever-type diseases which is well renowned as a mosquito-carried virus. Climate data is the primary cause of dengue epidemics, and it causes dengue which cases to fluctuate. The goal of this study is to deal with dengue forecasting using the machine learning method. The study contains 4 different machine learning models, those are- KNN, RF, GBR, and SVR. Before applying those models, all the data were preprocessed and in some cases, some of the attributes were deducted due to rare connections with the occurrence. Those connections were measured by correlation function. After all the models finalized the result it was time to compare those results depending on the mean absolute error(MAE). Through the whole procedure of the comparison and analysis, the result was denoting that the K-Nearest Neighbor(KNN) was giving significantly better performance than the other 3 models.

Keywords: Dengue, Machine Learning, Forecasting, Meteorological Data, Predictive Models.

1 Introduction

According to the Centers for Disease Control and Prevention, each year, up to 400 million people get infected with dengue and approximately 100 million people get sick from infection, and 40,000 die from severe dengue [1]. One-half of the world population is predicted to be affected in 2080 by dengue [2]. The virus that causes dengue fever is transmitted by the Aedes aegypti mosquito. Because dengue fever is spread by mosquitos, it is linked to weather and environmental factors such as temperature, precipitation, and vegetation. Increasing temperatures may worsen the problem by allowing dengue fever to spread and spread faster in low-risk or dengue-free areas of Asia, Europe, North America, and Australia [3]. To keep patients okay, supportive care is the main thing [4]. As a consequence, forecasting dengue epidemics is critical. With this forecast, health authorities throughout the world may take preventative actions to treat dengue fever before it spreads, potentially saving millions of lives.[5]

This paper aims to find the most efficient model for predicting dengue cases. So that it can prevent a dengue outbreak by accurately forecasting an increase in dengue cases. To achieve our goal we have utilized preprocessing techniques such as the mean method to fill the null data. Correlation also has been applied to select proper features which are more connected with the dengue incidence. Then we have used machine learning methods to predict dengue cases. One more thing we have added is we worked with weather data so the predictions depend on the weather.

The rest of the paper is organized as follows. Section 2 is all about related works which have been on the same topics. Section 3 contains dataset overview, section 4 contains the methodology, while section 5 has the results and analysis, followed by the conclusion and future work in section 6.

2 Related Works

Because of the seriousness of this disease, many approaches have been taken to forecast the dengue outbreak. In 2014, researchers found that land-use factors other than human settlements, such as different types of agricultural land, water bodies, and forests, can be linked to reported dengue cases in the Malaysian state of Selangor, using boosted regression to account for non-linearities and interactions between these factors [6]. In 2017, it was claimed that human migration has a massive effect on dengue pathogen transmission to immunologically dengue 'naive' regions [7]. Random forest was used to suggest a model for predicting dengue, diabetes, and swine flu in 2017. The major goal of this model is to forecast disease using patient symptoms and to propose a specialist doctor, as well as to calculate the dangerous instances of that particular disease in a week in that particular location [8]. P.Muhilthini et al introduced a dengue potential predicting model in 2018, and the data set includes information on the number of dengue cases recorded every week for several years in a variety of nations. It provides information on meteorological conditions such as temperature, precipitation quantity, humidity, and so on. GBR is used to find patterns and dependencies in the training data set and to predict the number of dengue cases for a specific week and year in the test data set [9]. With dengue, demographic, entomological, and environmental data in Singapore, Ong et al employed random forest regression to estimate the risk rank of dengue transmission in 1km grids. More than 80% of the observed risk rankings were within the predicted range of 80% [10]. In Thailand, Rachata et colleagues investigated the use of ANNs combined with an entropy approach to developing a prediction model for Dengue epidemics [11]. An ANN (Artificial Neural Network) was presented in research in Sri Lanka that used previous weather patterns and prior dengue cases as inputs to forecasting the dengue epidemic in the Kandy area [12].

3 Dataset Overview

We have taken data from the Driven Data[13] website with attributes Dengue Case and Meteorological dataset. There are 1456 records in the collection with 24 characteristics. Location, Temperature, Precipitation, Humidity, and Vegetation Index are the main features along with dengue total case. The dataset consists of two cities Sanjuan and Iquitos that derive the meteorological data that shows the mosquitoes are responsible for Dengue. This weekly-based dengue dataset is used for forecasting. Since the majority of the recordings are continuous, a regression model is constructed. Sanjuan's data was collected between 1999 and 2008 there are 936 records with 24 features, and it is one of the most affected cities in terms of dengue cases. Furthermore, the Iquitos data includes the years 2000 to 2010, with 520 recordings and 24 characteristics. It has a lesser effect than Dengue. We used the dataset for both the cities separately for our constructed model.

We developed a graphical representation of the total case across time to better analyze the dataset. We plotted it using the Python module. The first graph depicts the overall number of cases in Iquitos, while the second graph displays the number of cases in San Juan.

Table 1. details about dengue surveillance project data.

Timeline of Data	Total Records	Feature
1999-2008 (Sun Juan)	936	24
2000-2010 (Iquitos)	520	24

4 Methodology

Data mining is the process of studying and extracting information from large previous databases to predict unknown data on a particular example from previously observed examples. Here we used regression analysis to predict the dengue cases in a given situation from the provided dataset. We did this method before correlation with the whole dataset and then after finding a correlation, we omitted some features which have less correlated to the target feature and follow this method again to find the better result for discussing performance and comparative analysis.

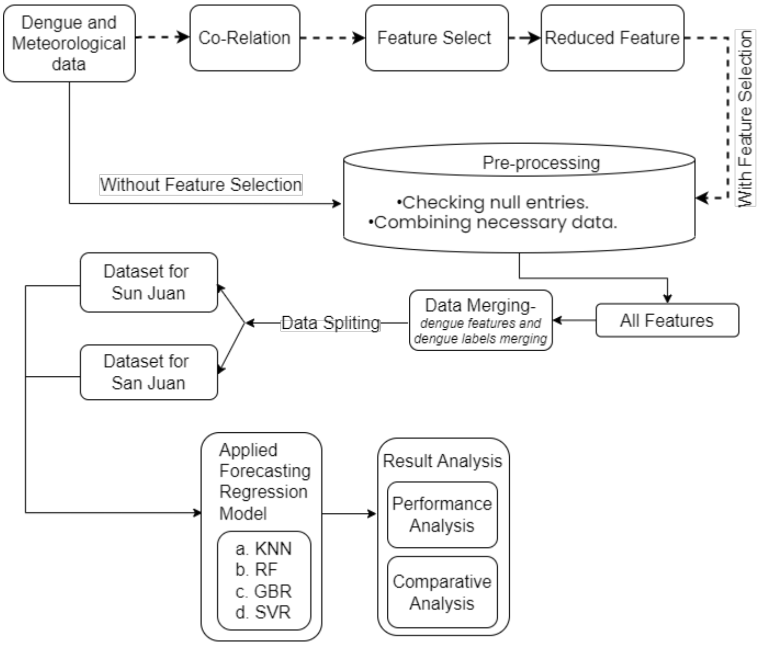

Fig. 1. Flow diagram using the feature selection method

In the first case of fig 1, we are giving the whole dataset into preprocessing where we check if any null entries are found then it is replaced with the mean value of that feature. After which convert all of the temperatures to Centigrade from Kelvin. Then rounding every value up to 3 decimal places. And combining necessary data as their average if needed. Afterward,s we merge dengue features with total dengue cases. As the dengue cases of Sun Juan and Iquitos are not dependent, so splitting them into different data-frames. And then we split the data-frames into train and test. After that, we apply four forecasting regression models (K Neighbors Regressor(KNN), Random Forest(RF), Gradient Boosting Regressor (GBR), Support Vector Regressor(SVR). Finally, we analyze the results of these models by using mean absolute error (MAE),

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{\text{n}} \qquad (1)$$

Equation (1) denotes the equation for generating the mean absolute error. Here, $y_i = prediction,$ $x_i = true\ value\ and$ n = total number of data points. In the second case, firstly we calculate correlations of each feature depending on the total dengue case then we optimize some feature which has a very low or negative correlation. After feature selection, we followed everything as the first case. After that, we compare and analyze the results to find out the best result and best model for this purpose.

4.1 Feature Selection

It is important to determine the correlations between the dataset attributes. As the ‘total case’ is our main target attribute which is denoting the total case of dengue that appeared in that specific timeline, we have to compare every single attribute with ‘total case’. We did the whole correlation part by excel statistical functions, it generally calculates the correlation coefficient between two variables. The formula is-

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \qquad (2)$$

After the calculation by equation (2), we deducted some of our attributes because those were not as much connected with the dengue occurrence as we wanted. So, in that case, we deducted all the negative values and some of the minimal positive values. As a result, in the case of San Juan from a total of 16 features, we keep 11 features for our experiment. On the other hand in the case of Iquitos from a total of 16 features, we keep 8 features for our experiment.

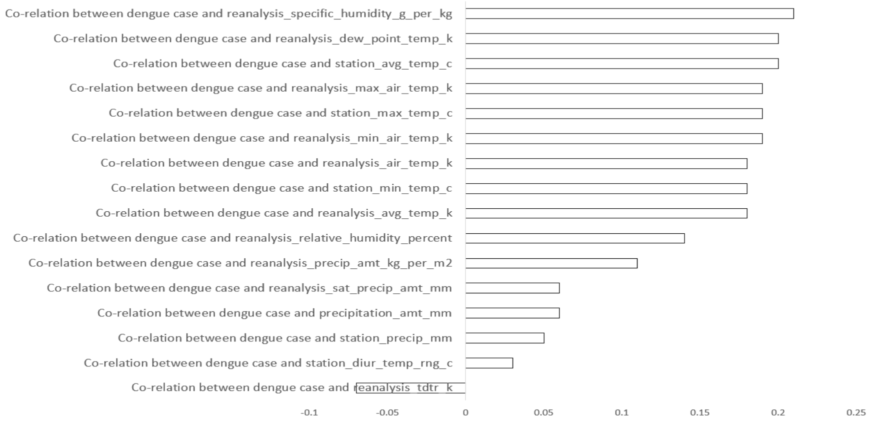

Fig. 2. Correlation between all the features vs ‘total case’ in San Juan.

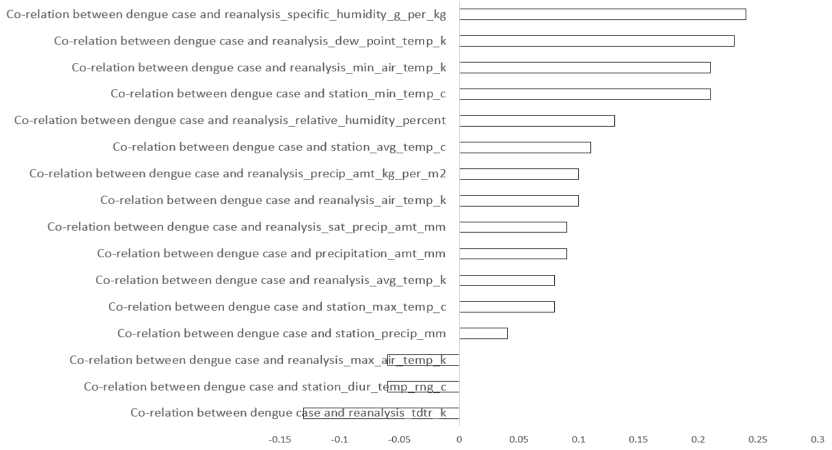

Fig. 3. Correlation between all the features vs 'total case' in Iquitos.

There are figure 2 and figure 3, we calculate the correlations between total cases with everything else to see what is more connected with dengue cases.

4.2 Data Preprocessing

We preprocessed our dataset in two ways. First, the data set before correlation was processed and then the dataset with correlated features was processed. The dataset had some missing values, if they had been eliminated then the authenticity of the dataset would have been reduced. There are a lot of techniques for dealing with missing data.

$$mean = \frac{sum\ of\ the\ terms}{numbers\ of\ terms} \qquad (3)$$

We choose to fill the missing dataset using a mean method using equation (3).

4.2 Model Description

K Neighbors Regressor(KNN)

The KNN algorithm predicts the values of new data points based on 'feature similarity.' This means that a value is assigned to the new point based on how similar it is to the points in the training set [14][15]. The essential assumption behind using kNN to analyze univariate time series is that consistent data-generating methods usually result in observations of repetitive patterns of activity[16][17].

Random Forest(RF)

Random Forests were invented by Leo Breiman, who was motivated by Amit and Geman's earlier work[18]. Random Forests are an outgrowth of Breiman's bagging concept[19] and were designed as a competitor to boosting, despite the description[20]. Random Forest is a "Tree"-based system that creates the final output by combining the results of numerous Decision Trees. They can be fine-tuned, but they frequently perform flawlessly with the default tuning parameters.[21]Fig 4 illustrates how the random forest algorithm works.

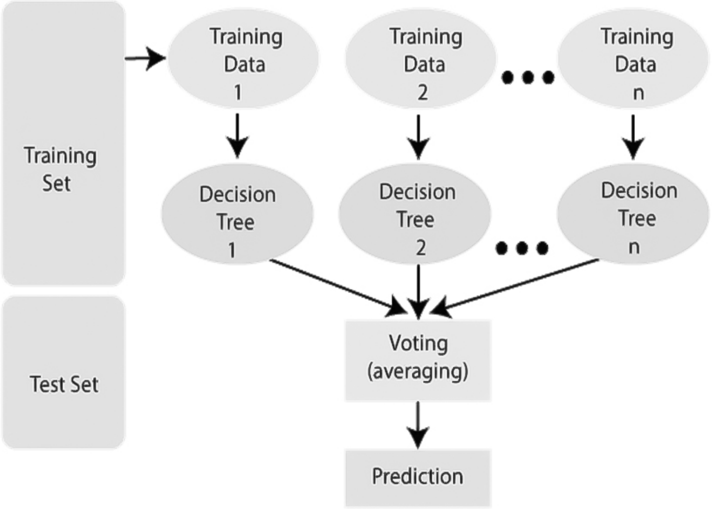

Fig. 4. A simplified illustration of a random forest technique.

Gradient Boosting Regressor (GBR)

Gradient Boosting Regressors generate an ensemble of shallow and weak successive trees with each tree learning and improving on preceding.

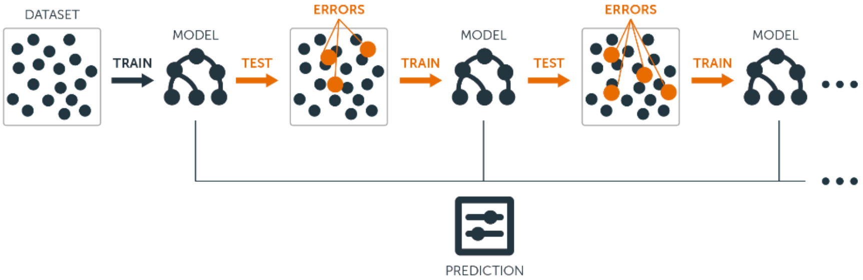

Fig. 5. Steps of Boosting technique.[22]

The primary idea behind boosting is to incrementally add additional models to the ensemble. Fig 5 illustrates how it works.

$$f(x) = \sum_{b=1}^{B} f^{b}(\mathrm{x}) \qquad (4)$$

Equation (4) is a derivation of the fundamental approach for boosted regression trees, where the final model is just a stagewise additive model of b individual regression trees[23]. The way a new distribution is built for the learning technique to produce the next hypothesis in the sequence, and the technique hypotheses are integrated to produce a highly accurate output, are two essential parts of boosting algorithms[24]. Both of these cases include a collection of parameters[25].

Support Vector Regressor(SVR)
Support-vector machines (also known as support-vector networks) are supervised learning models that examine data for classification and regression analysis in machine learning.[26] The goal of SVR is to fit the error within a particular threshold, which means that the goal of SVR is to approximate the best value within a given margin known as the ε- tube [27]. SVR uses an ε-insensitive loss function, which penalizes predictions that are too far away from the desired output. The width of the tube is determined by the value of ε [28].

5 Result

5.1 Result Analysis

To achieve the best results, we used four machine learning regression model methods in this paper. First, we used the KNN regressor model on a preprocessed dataset, and we got a mean absolute error(MAE) of 19.90 for Sanjuan and 5.52 for Iquitos. For the random forest model, we received MAE for Sanjuan 20.56 and Iquitos 5.41 in the second trial. After that, we used the Gradient Boosting Regressor to get MAE for Sanjuan, which is 21.69, and Iquitos 5.39. Finally, we used the Support Vector Regressor, which resulted in a value of 26.70 for San Juan and 4.03 for Iquitos. Before feature selection, we may say that KNN provided better results for Sanjuan, while SVR produced better results for Iquitos. After that, we used the correlation approach to select features before applying models and calculating mean absolute error for both cities. After feature selection, the result was significantly better than before. KNN is Sanjuan's best model, with an MAE of 16.88, and SVR is Iquitos' best model, with an MAE of 4.54. Sanjuan's MAE was 1.34, 17.14, and 19.92, correspondingly, using Random forest, Gradient boosting, and Support Vector Regression models. MAE values for KNN, RF, and GBR were 5.14,5.21, and 5.58, consecutively, for Iquitos. For this study, we used the total results for both cities where KNN proved to be the best model.

Table 2. The comparison between the MAE before and after feature selection.

Model	Before Feature Selection		After Feature Selection	
	San Juan	Iquitos	San Juan	Iquitos
KNN	19.90	5.52	16.88	5.14
RF	20.56	5.41	17.34	5.21
GBR	21.69	5.39	17.14	5.58
SVR	26.70	4.03	19.92	4.54

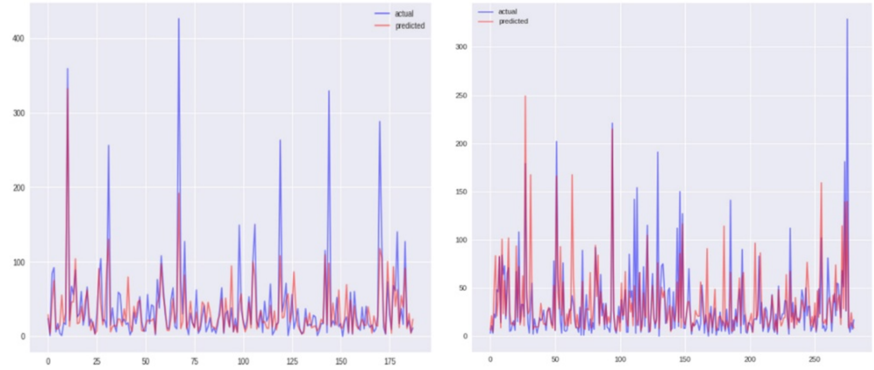

Fig. 6. Comparison between true values and predicted values using KNN model before and after correlation (Sun Juan)

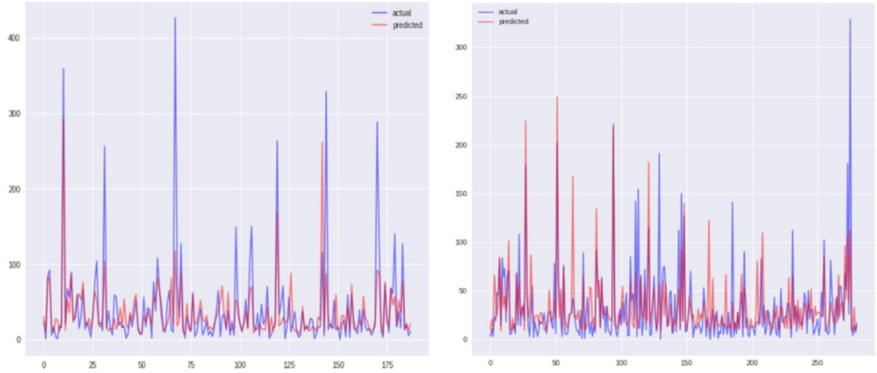

Fig. 7. Comparison between true values and predicted values using RF model before and after correlation (Sun Juan)

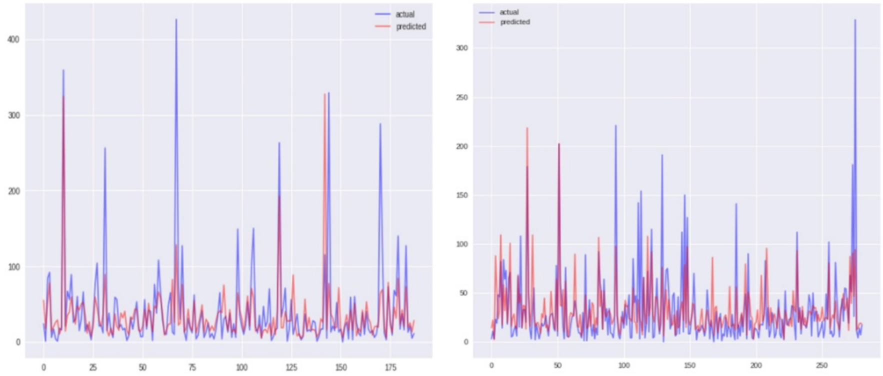

Fig. 8. Comparison between true values and predicted values using GBR model before and after correlation (Sun Juan)

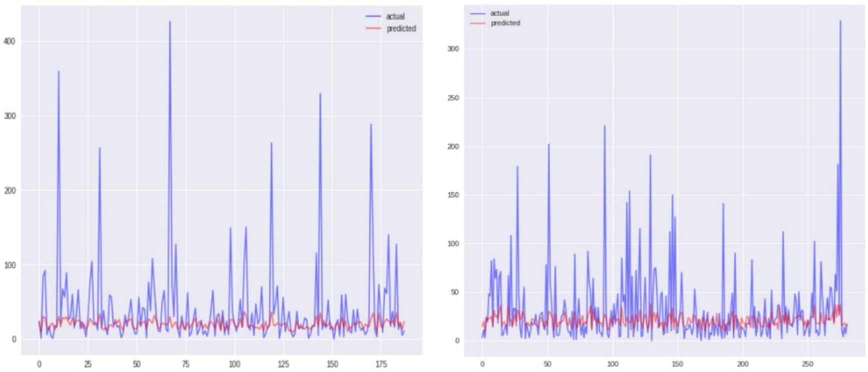

Fig. 9. Comparison between true values and predicted values using SVR model before and after correlation (Sun Juan)

For each of the four models, we can see a graphical representation of Sanjuan city in Figures 6, 7, 8, and 9 showing the actual and predicted values. However, figure 6 shows that the values are closer together than other figures, indicating that KNN is the best model for this.

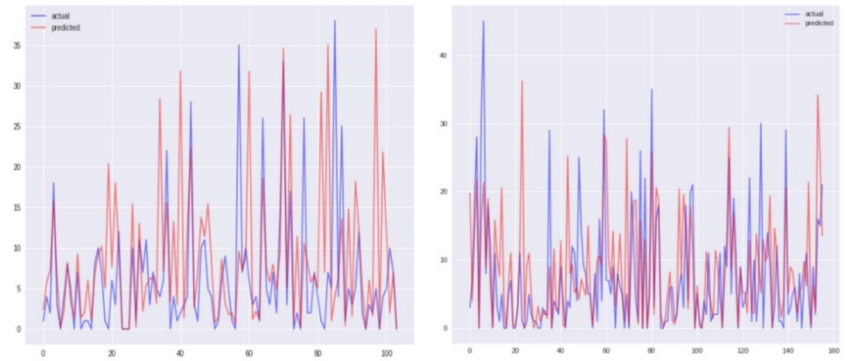

Fig. 10. Comparison between true values and predicted values using KNN model before and after correlation (Iquitos)

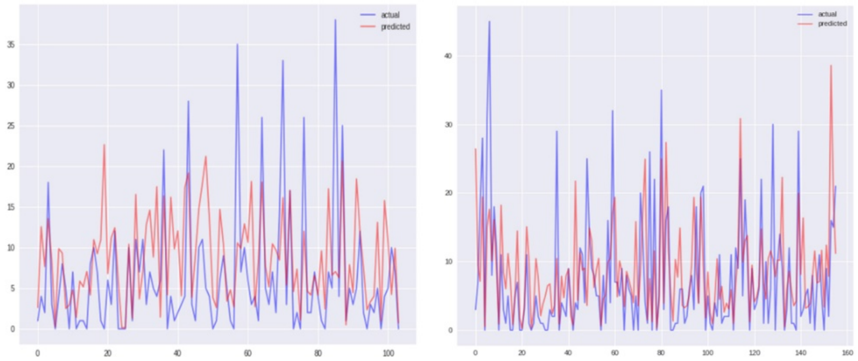

Fig. 11. Comparison between true values and predicted values using RF model before and after correlation (Iquitos)

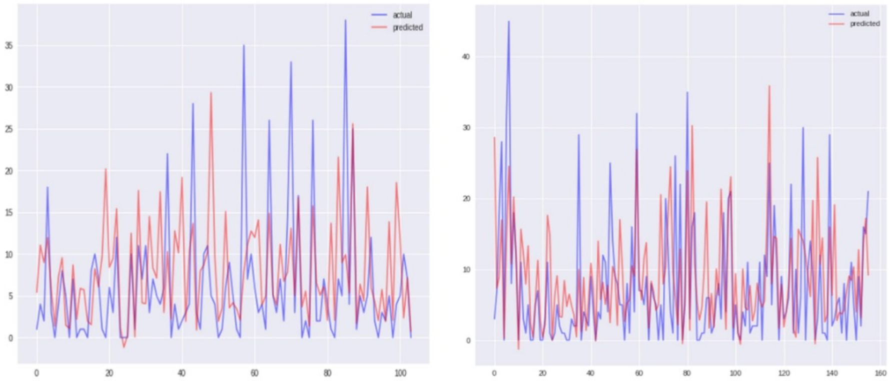

Fig. 12. Comparison between true values and predicted values using GBR model before and after correlation (Iquitos)

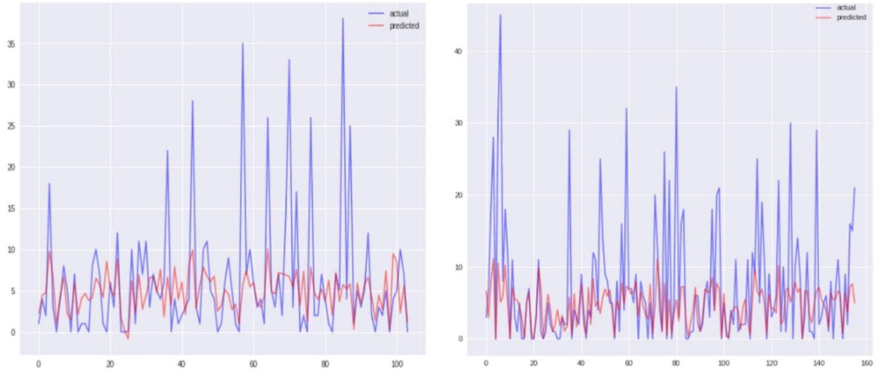

Fig. 13. Comparison between true values and predicted values using SVR model before and after correlation (Iquitos)

Figures 10, 11, 12, and 13 show a graphical representation of Iquitos city for each of the four models, as well as the actual and expected values. Figure 13 shows that the values are closer together in this figure than in the others, indicating that SVR is the best model for the city.

5.2 Result Comparison

We have compared our results to previous work on the same dataset[29]. we can conclude that, in comparison to past efforts, we created a decent outcome, and that we also excelled in implementing the new model and achieving greater results.

Table 3. The comparison between the MAE before and after feature selection.

Model	This Paper		[29]**	
	San Juan	Iquitos	San Juan	Iquitos
KNN*	16.88	5.14	-	-
RF	17.34	5.21	26.66	6.70
GBR	17.14	5.58	24.11	25.98
SVR*	19.92	4.54	-	-

* These models are not used by the referenced paper we compared with.

** The publication with which we've made a comparison.

6 Conclusion

We worked on a continuous data set. A continuous data set is a quantitative data set representing a scale of measurement that can consist of numbers other than whole numbers, like decimals and fractions. Continuous data sets would consist of values like height, weight, length, temperature, and other measurements like that[30]. Thus we couldn’t apply any classification model.

Dengue is a disease that affects a wide range of people. This paper demonstrated a machine learning-based intelligent system capable of forecasting dengue incidence based on weather variability using real-world data. The findings indicate that variations in dengue incidence are linked to climate variability conditions for the two selected cities. Furthermore, the main climate element that influences dengue incidence differs in different ways from one city to the next. We achieved a minimum mean absolute with the major models, indicating that a good forecasting model is built and KNN performed the best out of all the models.

References

1. Centers for Disease Control and Prevention, “About Dengue: What You Need to Know,” [Online]. Available: https://www.cdc.gov/dengue/about/index.html#:~:text=Almost half of the world’s,40%2C000 die from severe dengue.
2. J. P. Messina *et al.*, “The current and future global distribution and population at risk of dengue,” *Nat. Microbiol.*, vol. 4, no. 9, pp. 1508–1515, 2019, doi: 10.1038/s41564-019-0476-8.
3. T. P. Monath, “Dengue: the risk to developed and developing countries.,” *Proc. Natl. Acad. Sci.*, vol. 91, no. 7, pp. 2395–2400, 1994, doi: 10.1073/pnas.91.7.2395.
4. D. Wallace, V. Canouet, P. Garbes, and T. A. Wartel, “Challenges in the clinical development of a dengue vaccine.,” *Curr. Opin. Virol.*, vol. 3, no. 3, pp. 352–356, Jun. 2013, doi: 10.1016/j.coviro.2013.05.014.
5. cdc, “Stages of Dengue,” [Online]. Available: https://www.cdc.gov/dengue/training/cme/ccm/page86100.html.
6. Y. L. Cheong, P. J. Leitão, and T. Lakes, “Assessment of land use factors associated with dengue cases in Malaysia using Boosted Regression Trees.,” *Spat. Spatiotemporal. Epidemiol.*, vol. 10, pp. 75–84, Jul. 2014, doi: 10.1016/j.sste.2014.05.002.
7. K. G. S. Dharmawardana *et al.*, “Predictive model for the dengue incidences in Sri Lanka using mobile network big data,” in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2017, pp. 1–6, doi: 10.1109/ICIINFS.2017.8300381.
8. A. Tate, U. Gavhane, J. Pawar, B. Rajpurohit, and G. B. Deshmukh, “Prediction of Dengue, Diabetes and Swine Flu Using Random Forest Classification Algorithm,” 2017.
9. P. Muhilthini, B. Meenakshi, S. Lekha, and S. Santhanalakshmi, “Dengue Possibility Forecasting Model using Machine Learning Algorithms,” 2018.
10. J. Ong *et al.*, “Mapping dengue risk in Singapore using Random Forest,” *PLoS Negl. Trop. Dis.*, vol. 12, no. 6, pp. 1–12, 2018, doi: 10.1371/journal.pntd.0006587.
11. N. Rachata, P. Charoenkwan, T. Yooyativong, K. Chamnongthal, C. Lursinsap, and K. Higuchi, “Automatic Prediction System of Dengue Haemorrhagic-Fever Outbreak Risk by Using Entropy and Artificial Neural Network,” in *2008 International Symposium on Communications and Information Technologies*, 2008, pp. 210–214, doi: 10.1109/ISCIT.2008.4700184.
12. H. Wijekoon, P. Herath, and A. Perera, “Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks,” *Int. J. Comput. Appl.*, vol. 101, p. 1, 2014, doi: 10.5120/17760-8862.
13. DengAI: Predicting Disease Spread", *DrivenData*, 2022. [Online]. Available: https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/.

14. Y. Song, J. Liang, J. Lu and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression", *Neurocomputing*, vol. 251, pp. 26-34, 2017. Available: 10.1016/j.neucom.2017.04.018
15. Ban, T., Zhang, R., Pang, S., Sarrafzadeh, A. and Inoue, D., 2013. Referential kNN Regression for Financial Time Series Forecasting. *Neural Information Processing*, pp.601-608.
16. Meade, N., 2002. A comparison of the accuracy of short term foreign exchange forecasting methods. *International Journal of Forecasting*, 18(1), pp.67-83.
17. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G., Ng, A., Liu, B., Yu, P., Zhou, Z., Steinbach, M., Hand, D. and Steinberg, D., 2007. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1), pp.1-37.
18. Amit, Y. and Geman, D., 1997. Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, 9(7), pp.1545-1588.
19. L. Breiman, *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996. Available: https://link.springer.com/content/pdf/10.1023/A:1018054314350.pdf.
20. Breiman, L.: Random Forests. Machine Learning 45 (1) pp. 5–32 (2001).
21. A. Cutler, D. Cutler and J. Stevens, "Random Forests", *Ensemble Machine Learning*, pp. 157-175, 2012. Available: 10.1007/978-1-4419-9326-7_5 .
22. "Understand different types of Boosting Algorithms", *OpenGenus IQ: Computing Expertise & Legacy*, 2022. [Online]. Available: https://iq.opengenus.org/types-of-boosting-algorithms/.
23. "Gradient Boosting Machines · UC Business Analytics R Programming Guide", *Uc-r.github.io*, 2022. [Online]. Available: https://uc-r.github.io/gbm_regression.
24. A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial", *Frontiers in Neurorobotics*, vol. 7, 2013. Available: 10.3389/fnbot.2013.00021
25. "A Gradient-Based Boosting Algorithm for Regression Problems", *Proceedings.neurips.cc*, 2022. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/8d9fc2308c8f28d2a7d2f6f48801c705-Paper.pdf.
26. Chen-Chia Chuang, Shun-Feng Su, Jin-Tsong Jeng and Chih-Ching Hsiao, "Robustsupport vector regression networks for function approximation with outliers", *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1322-1330, 2002. Available: 10.1109/tnn.2002.804227 .
27. S. Regression, "Support Vector Regression | Learn the Working and Advantages of SVR", *EDUCBA*, 2022. [Online]. Available: https://www.educba.com/support-vector-regression/.
28. M. Awad and R. Khanna, "Support Vector Regression", *Efficient Learning Machines*, pp. 67-80, 2015. Available: 10.1007/978-1-4302-5990-9_4
29. Singh, Chanpreet & Anuranjan,. (2019). Predicting Dengue Spread in San Juan and Iquitos using Machine Learning. 10.13140/RG.2.2.24207.74406.
30. A. Zangre, "Discrete vs Continuous Data – What's the Difference?", *https://www.g2.com/*, 2022. [Online]. Available: https://www.g2.com/articles/discrete-vs-continuous-data.