

How to Cite:

Chandra, D., Islam, S., & Pandey, D. (2022). Dengue prediction using time series forecasting I.E ARIMA model. *International Journal of Health Sciences*, 6(S8), 4517–4526. <https://doi.org/10.53730/ijhs.v6nS8.13221>

Dengue prediction using time series forecasting I.E ARIMA model

Dinesh Chandra

Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, India

Corresponding author email: dinmax007@gmail.com

Shaziya Islam

Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, India

Email: shaziya.islam26@gmail.com

Deepak Pandey

Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai, India

Email: 14deepak1993@gmail.com

Abstract---Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. It is a life threatening disease lots of person died due to dengue because its symptoms are not detected at early stages many parson thought that it was a normal fever or headache so that they just ignore it which cause there are in danger situations and in worst case they loss their life.

Keywords---ARIMA model, time series forecasting, component.

Introduction

Dengue is a life intimidating disease prevalent in several developed also as developing countries like India. This is often an epidemic born disease caused by breeding of Aedes mosquito. Datasets that are available for dengue describe information about the patients suffering with dengue disease and without dengue disease alongside their symptoms like: Fever Temperature, WBC, Platelets, Severe Headache, Vomiting, Metallic Taste, Joint Pain, Appetite, Diarrhea, Hematocrit, Hemoglobin, and the way many days suffer in several city.[1][2][3] Although dengue has been known for more than 200 years, it was only in 1950 that the first dengue viruses were isolated. The incidence of dengue infection has increased approximately 30 times since 1950. According to the World Health Organization (WHO), and mentioned by Stanaway et al. about 20,000 people per

year die from this disease. Since 1970, epidemics of dengue infection have occurred in more than 128 countries worldwide with a significant expansion of the spread of the disease to new areas of the world.

Dengue viruses often have a relationship with other mosquito-spread viruses, such as Yellow Fever, Zika, Japanese Encephalitis, and West Nile viruses. Dengue virus is classified into 4 serotypes: DEN-1, DEN-2, DEN-3, and DEN-4. lots of person died due to dengue because its symptoms are not detected at early stages many parson thought that it was a normal fever or headache so that they just ignore it which cause there are in danger situations and in worst case they loss their life.

In this system we applied data mining techniques along with arima model to create a system which can predict dengue with the help of symptoms provided by the users in dataset. In data mining portion we use data filtering, data cleaning and clustering and some another techniques to enhance our dataset. ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data. ARIMA stands for the Autoregressive integrated moving average model, that is one of the techniques used in the time series analysis for forecasting. There are styles of ARIMA approach; the non-seasonal Autoregressive integrated moving average and the seasonal Autoregressive integrated moving average (SARIMA)[4]

Literature review

In In 2018 P. Siriyasatien¹, S. Chadsuthi², K. Jampachaisri³, K. Kesorn^{*,4} proposed a Dengue Epidemics Prediction: A Survey of the State-of-the-Art based on Data Science Processes where analyzes the major components that can be used in a dengue prediction model. They have attempted to identify the factors directly related to the probability of a dengue epidemic occurring, particularly climate factors, the rate of mosquito bites, rainfall, and the rate of dengue infection in mosquitoes as the important factors contributing to severe outbreaks. Another surveillance method for dengue outbreaks surveillance is to ascertain information posted by tourists on social networks. The immediacy of this information quickly alerts the international community about the prevalence of the disease in an area. Dengue outbreaks are a risk for all travelers who should, therefore, study the seasonal risk factors at their intended destinations, and seek and receive accurate advice about preventive measures during their travel [5].

In 2018 Abdul Mahatir Najar, Nopember Surabaya, Mohammad Isa Irawan proposed an Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction where they apply extreme learning machine (ELM) method to predict the risk of outbreak based on weather condition. They Develop ELM architecture with weather variables as input nodes and risk level of DHF outbreak as the target. They use binary sigmoid activation function and bipolar sigmoid with a number of hidden neurons between 5 - 200 nodes. The results show that ELM can predict the level of risk of DHF with the best performance of ELM network using a binary sigmoid activation function with 50 hidden neurons [6].

In 2018 Virginia Ortiz Andersson, Marco A. Ferreira Bircky and Ricardo Matsumura Araujo proposed a methodology on Towards Predicting Dengue Fever Rates Using Convolutional Neural Networks and Street-Level Images where they investigate the use of street-level images, such as those from Google Street View, along with Convolutional Neural Networks to predict Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) rates in urban locations. They conduct a case study in the city of Rio De Janeiro, Brazil, using the proposed methodology and DF/DHF data between the years of 2010 to 2014. We compare two Siamese based CNNs, yielding an overall accuracy of 67% for 20,400 different locations. We conclude that street-level images are useful for the problem [7].

In 2018 Ria Arafiah¹ and Fariani Hermin¹ proposed a Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method where The use of data mining method, able to build data base support in decision makers diagnose DHF disease This study predicts DHF with the method of Naive Bayes. Parameter of The input variable is the patient's medical data (temperature, spotting, bleeding, and tornuine test) and the output variable suffers from DBD or not while the system output is diagnosis of the patient suffering from DHF or not. Result of model test by using tools of Orange 3.4.5 obtained level of precision model is 77,3% [8].

In 2018 P. Sathya [1], Mrs. A.Sumathi [2] proposed a method for Predicting Dengue Fever Using Data Mining Techniques where they are trying to find out some of characteristics on Dengue illness so that they can rightly categorize patients because different patients require different types of treatment. Pakistan has been target of Dengue illness from last few years. Dengue fever is used in clustering techniques to evaluate their act. The dataset was gathered from lotus and 24 care hospitals. For properly classify our dataset, various clustering techniques are used. Appraise the performance of all the techniques individual based on tables and graphs depending upon dataset [9].

In 2017 Kashish are Shakil, Shadma Anis and Mansaf Alam proposed a DENGUE Disease Prediction Using WEKA DATA MINING TOOL where they have firstly classified the dengue data set and then compared the different data mining techniques in weka through Explorer, knowledge flow and experimenter interfaces. Furthermore in order to validate our approach we have used a dengue dataset with 108 instances but weka used 99 rows and 18 attributes to determine the prediction of disease and their accuracy using classifications of different algorithms to find out the best performance. Their main objective is to classify data and assist the users in extracting useful information from data and easily identify a suitable algorithm for accurate predictive model from it. From the findings of this paper it can be concluded that Naïve Bayes and J48 are the best performance algorithms for classified accuracy because they achieved maximum accuracy= 100% with 99 correctly classified instances, maximum ROC = 1 , had least mean absolute error and it took minimum time for building this model through Explorer and Knowledge flow results[10].

In 2017 H. Abdul Rahim¹, F. Ibrahim proposed a novel prediction system in dengue fever using narmax model where the development of nonlinear autoregressive moving average with exogenous input (NARMAX) models in

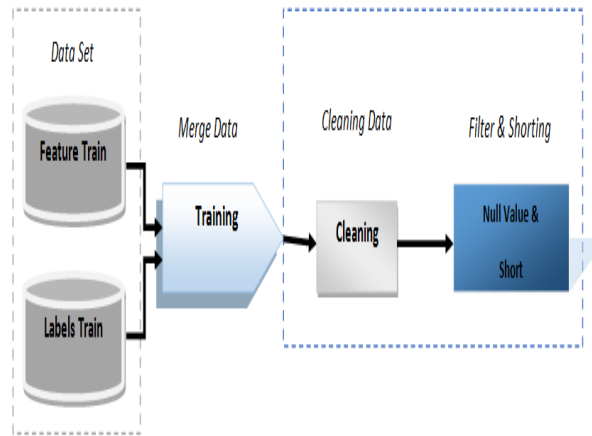
diagnosing dengue infection. The developed system bases its prediction solely on the bioelectrical impedance parameters and physiological data. Three different NARMAX model order selection criteria namely FPE, AIC and Lipschitz have been evaluated and analyzed. This model is divided into two approaches which are unregularized approach and regularized approach. The results show that using Lipschitz number with regularized approach yield better accuracy by 88.40% to diagnose the dengue infections disease. Furthermore, this analysis shows that the NARMAX model yield better accuracy as compared to autoregressive moving average with exogenous input (ARMAX) model in diagnosis intelligent system based on the input variables namely gender, weight, vomiting, reactance and the day of the fever as recommended by the outcomes of statistical tests with 76.70% accuracy [11].

In 2017 Ms.S.Freeda Jebamalar, Dr.A.Anitha proposed A Survey on Prediction of Dengue Fever Using Data Mining Techniques where they survey on dengue disease and its symptoms where they survey the importance of data mining techniques and they concluded that data mining techniques are more effective in predicting the dengue disease with accuracy ranging from 70% to 100% [12].

In 2017 P. Manivannan, Dr. P. Isakki Devi proposed a method for Predicting Dengue Fever Prediction Using K-Means Clustering Algorithm where has been focused four stages namely preprocessing, attribute selection, clustering and predicting the dengue fever. R 3.3.2 Tool is used for preprocessing the household of dengue dataset. D win's method has been applied to generate filled dataset by substituting all missing values for nominal and numeric attributes with mode and mean value. Dengue virus can be predicted by applying different data mining techniques. The main goal of research work is to predict the people who are affected by dengue depending upon categorization of age group using K-means clustering algorithm has been implemented [13].

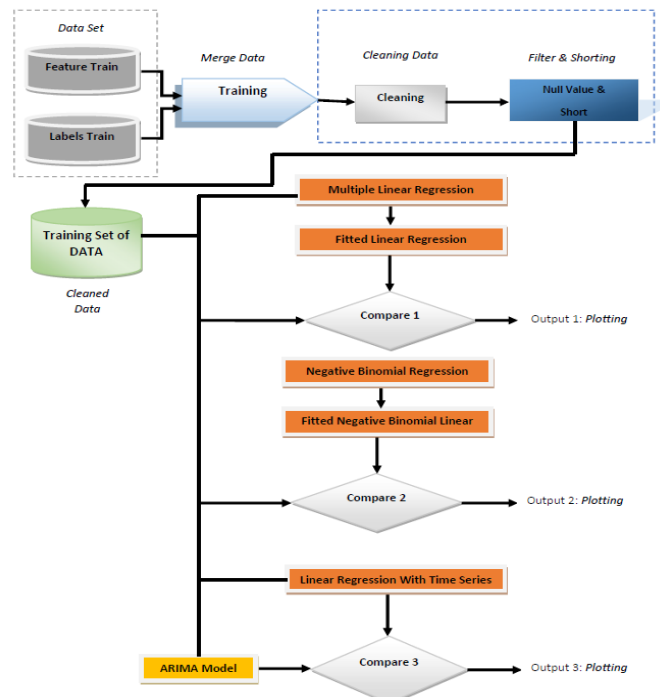
Methodology

Here we using two Data set first is weather data set which is used for training and our system. And another Data set is used for Dengue prediction, which contains details of Dengue patients. In our Weather data set, it contains various parameters Such as Average air temperatures, NDVI, minimum and maximum temperature, air moisture and certain outer parameter. In this project, we propose a methodology which used two data set one is for training and the other is used for testing. Our training data set is based on records of dengue patients and their sickness symptoms for creation of Data set we use two parameters i.e. feature term and label train. Now, we merge feature train and labels train to create a training data set. Then we apply Data mining technique to extract our data such as cleaning, filtering and shortness. Now our training Data set is created.



Algorithm: Dengue Prediction

- 1) Creation of Data set which Contains dengue cases of Rajasthan and Chhattisgarh.
- 2) Applying Data Mining Techniques.
 - a) Load data from data set.
 - b) Then we apply data cleaning so that detecting and removing corrupt or inaccurate records from a data set. As a result str feature train and str label train is created.
 - c) Merging features train with label train data.
 - d) Checking Summery for training set where we removing unknown columns so that our processing task becomes fast.
 - e) Checking Null values then replacing null values by mean values.
- 3) Then we apply Data visualization where we show graphical representation of information and data. By using visual elements like charts, graphs, and maps. In this phase we perform following techniques :-
 - a) See the correlation between numeric variable
 - b) Visualizing using the correlation plot.
 - c) Look for correlation as bar plot.
 - d) Show correlation bar plot in grid.
- 4) Then we applied predictive analysis In this phase we applied following techniques :-
 - a) Multiple linear regressions.
 - b) Negative binominal regression.
 - c) Linear regression with time series.
- 5) Now we applied ARIMA model which perform following operation :-
 - a) Plotting actual vs Predicted Data result.
 - b) Check residuals
 - c) Compare ARIMA result with linear regression.
- 6) Result in the form of graph is shown which predicts dengue results.



Result

ARIMA stands for the Autoregressive integrated moving average model that is one of the techniques used in the time series analysis for forecasting. There are styles of ARIMA approach; the non-seasonal Autoregressive integrated moving average and the seasonal Autoregressive integrated moving average (SARIMA). Time series algorithms are used extensively for analyzing and forecasting time-based data. One set of popular and powerful time series algorithms is the ARIMA class of models, which are based on describing autocorrelations in the data. ARIMA stands for *Autoregressive Integrated Moving Average* and has three components, p , d , and q , that are required to build the ARIMA model. These three components are:

p : Number of autoregressive lags

d : Order of differencing required to make the series stationary

q : Number of moving average lags

The ARIMA model was generated in this study to predict the outbreak of dengue. This model has been identified as one of the most effective methods used to predict time-series data. This method is divided into three phases, which are identification, estimation and forecasting phases. The actual data have to be stationary. If the data is not stationary, proceed to the identification stage.

Identification Phase:

Step 1: Evaluate the stationary data series with three approaches.

i) Plot time series plot for the actual data and check if a trend component appears on the plot.

ii) Plot the autocorrelation function (ACF) correlogram and partial autocorrelation function (PACF) correlogram. If decay occurs in the ACF plot, the stationary ACF plot can be determined by applying the Augmented Dicky-Fuller (ADF) test.

Decision Rule:

Reject H_0 if p-value ≥ 0.05

H_0 : The data is stationary.

H_1 : The data is not stationary.

iii) Check the PACF plot if there is a higher spike shown in the plot then the data is not stationary.

Step 2: Differencing when the data is non-stationary.

Differencing can also be checked using the Kwiatkowski–Phillips–Schmidt–Shin test (KPSS Test) whether or not differentiation is required. If the test statistics shown are greater than 10% of the critical value, the null hypothesis is rejected. Consequently, the data is not stationary. Repeat the data differencing and use the KPSS test again.

Decision Rule:

Reject H_0 if p-value > 0.1

H_0 : The data is stationary.

H_1 : The data is not stationary.

Estimation Phase:

Step 3: Check the parameters estimated for the ARIMA model.

The ARIMA model is categorized into three terms, p, d, and q, where p is the order of the autoregressive (AR) term, q is the order of the moving average (MA) term, and d is the number of differencing used to obtain the stationary time series data.

These are the mathematical

formulas for each term:

Autoregressive (AR), p term,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (1)$$

Moving Average (MA), q term,

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (2)$$

Where ε_t represents white noise.

Number of differencing, d term,

$$\Delta^d y_t = (1 - B)^d y_t \quad (3)$$

Where B is the lag operator.

Therefore, the general formula for the ARIMA model is as follows,

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (4)$$

Where y'_t is a differenced series.

Forecast Phase:

Step 4: Check the accuracy of the model to select the best ARIMA model. To verify the accuracy of the model, measuring errors of the ARIMA model, such as MSE, RMSE, and MAPE, are then calculated. Then, the best model is selected to forecast the data. In this model we compare our model with actual data using time series arima model

```

> #fit ARIMA model
> rj.model4 <- auto.arima(ts_rj[,1], xreg=ts_rj[,2:3])
> summary(rj.model4)
Series: ts_rj[, 1]
Regression with ARIMA(0,1,1) errors

Coefficients:
          ma1 specific_humidity_g_per_kg_reanalysis dew_point_temp_k_reanalysis max_temp_c_station
            -0.5602                1.1859                1.0086                -0.6475
s.e.         0.1381                13.7313                13.4151                0.9014
s.e.         0.9782
s.e.         2.1257

sigma^2 estimated as 36.46: log likelihood=-151.97
AIC=315.93 AICc=317.98 BIC=327.16

Training set error measures:
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 0.1750984 5.656632 4.144923 -inf  inf  nan 0.06621814
>
> cg.model4 <- auto.arima(ts_cg[,1], xreg=ts_cg[,2:3])
> summary(cg.model4)
Series: ts_cg[, 1]
Regression with ARIMA(2,0,0) errors

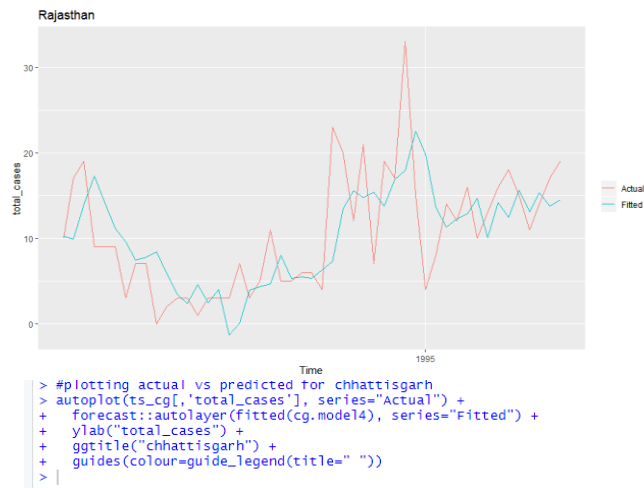
Coefficients:
          ar1      ar2 specific_humidity_g_per_kg_reanalysis dew_point_temp_k_reanalysis
            0.9247  -0.3528                -0.3763                0.0434
s.e.         0.1311  0.1301                0.5963                0.0361

sigma^2 estimated as 17.4: log likelihood=-140.72
AIC=291.45 AICc=292.81 BIC=301.01

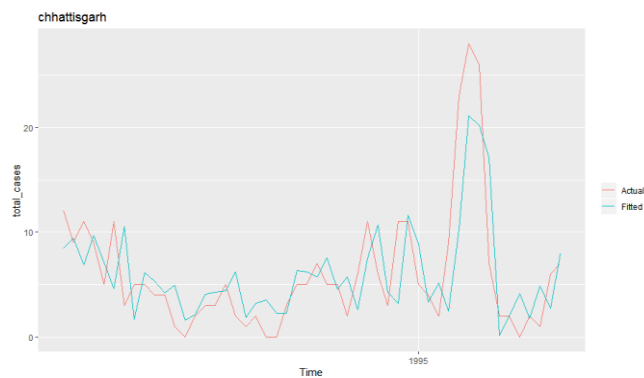
Training set error measures:
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set -0.04426262 4.001166 2.96722 -inf  inf  nan -0.0160309
~

```

Result as:



Result as:



```

> #check residuals
> checkresiduals(rj.model4)

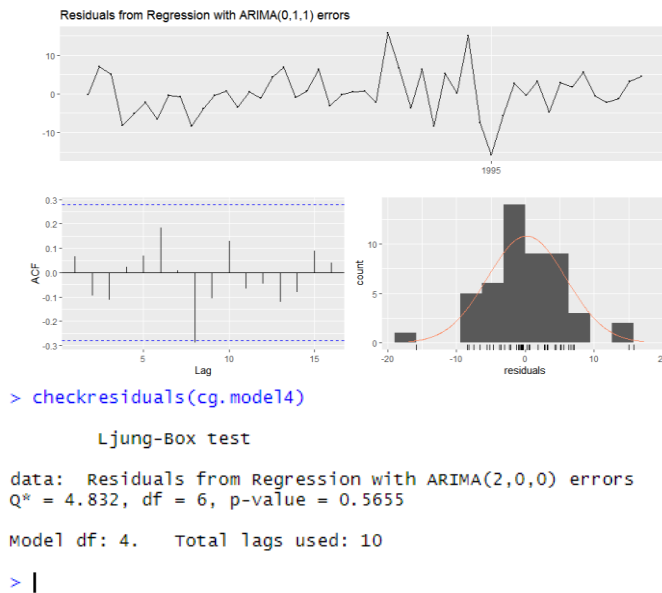
```

Ljung-Box test

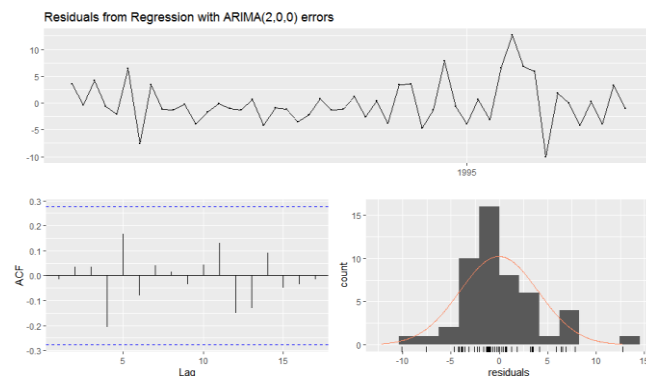
data: Residuals from Regression with ARIMA(0,1,1) errors
 $Q^* = 10.542$, $df = 5$, $p\text{-value} = 0.06125$

Model df: 5. Total lags used: 10

Result as:



Result as:



References

1. Abdul Rahim, H., Ibrahim, F., & Taib, M. N. (2007). A novel prediction system in dengue fever using NARMAX model. 2007 International Conference on Control, Automation and Systems. doi:10.1109/iccas.2007.4406927.
2. Andersson, V. O., Ferreira Birck, M. A., & Araujo, R. M. (2018). Towards Predicting Dengue Fever Rates Using Convolutional Neural Networks and Street-Level Images. 2018 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn.2018.8489567
3. Arafiah, R., & Hermin, F. (2018). Data mining for dengue hemorrhagic fever (DHF) prediction with naive Bayes method. Journal of Physics: Conference Series, 948, 012077. doi:10.1088/1742-6596/948/1/012077 .

4. Chengathir, S. M., Bhuvaneswari, T., Maruthupandi, J., & Naga, P. R. (2022). Prediction of dengue using data mining classification algorithms. *International Journal of Health Sciences*, 6(S1), 11860² 11871. <https://doi.org/10.53730/ijhs.v6nS1.7907>
5. Dr.J.VijiPriya et.al , "A Study on Development of Multilingual Dictionary Software.", *International Journal of Computer Engineering In Research Trends*, 4(9):pp:367-372 ,September-2017.
6. kashish, Anis, Shadma, Alam, Mansaf Ara Shakil, "Dengue Disease prediction Using Weka Data Mining tool," Elsevier, p. 26, Feb 2015.
7. Kishalay De , Ilya Mereminskiy,Roberto Soria,Charlie Conroy," A new Galactic symbiotic X-ray binary outburst triggered by an intense mass loss episode of a heavily obscured Mira variable," Draft version May 20, 2022 Typeset using LATEX twocolumn style in AASTeX631.
8. Kumar, R., Kumar, P., & Kumar, Y. (2021). Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms. *International Journal of Information Technology*. doi:10.1007/s41870-021-00741-8
9. Manivannan, P., & Devi, P. I. (2017). Dengue fever prediction using K-means clustering algorithm. 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). doi:10.1109/itcosp.2017.8303126 .
10. Mrs. A.Sumathi [2] P. Sathya [1], "Predicting Dengue Fever Using Data Mining Techniques," *IJCST*, vol. 6, no. 2, p. 3, March-April3 2018.
11. Ms.S.Freeda Jebamalar, Dr.A.Anitha , "A Survey on Prediction of Dengue Fever Using Data Mining Techniques," *IJSEM*, vol. 2, no. 12, p. 3, Dec 2017.
12. Najar, A. M., Irawan, M. I., & Adzkiya, D. (2018). Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction. 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE). doi:10.1109/icscee.2018.8538409
13. Siriyasatien, P., Chadsuthi, S., Jampachaisri, K., & Kesorn, K. (2018). Dengue Epidemics Prediction: A Survey of the State-of-the-Art based on Data Science Processes. *IEEE Access*, 1–1. doi:10.1109/access.2018.2871241