International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Dengue Outbreak Prediction using Regression Model in Chittoor District, Andhra Pradesh, India

J. Avanija, G. Sunitha, R. Hitesh Sai Vittal

Abstract*: Dengue fever is one of the tropical diseases, also called as break bone fever. It is caused by transmission of dengue virus through bite from Aedes mosquitoes. Dengue fever is previously predicted regionally using gradient descent algorithm. Annually 50 million to 528 million people get affected by dengue fever and 10,000 to 20,000 die. Reason behind this paper is to predict outbreak of dengue fever regionally by using features such as temperature, rainfall, humidity because breeding of Aedes mosquitoes is related to these features. Bagging and boosting techniques are used with gradient descent to predict more accurately the occurrence of dengue in a region. Data is collected from regional government weather office. Data is pre-processed like filling missing values and normalizing values. Feature selection processes like dimensionality reduction is done on pre-processed dataset. Gradient descent is applied with bagging and boosting. Accuracy is calculated by plotting graphs and by calculating Mean Standard Deviation (MSD) and Mean Absolute Error (MAE). By prediction of dengue fever before its occurrence in a region, makes it easy to vaccinate people in that region and dengue can be controlled and stopped from becoming an outbreak.

Keywords*: Epidemic Outbreak, Data Analysis, Prediction Modeling, Deep Learning.

I. INTRODUCTION

Dengue is a tropical disease which is mosquito borne caused by a single-stranded RNA virus belonging to the family Flaviviridae of genus flavivirus. Five serotypes of virus have been found, which can cause full spectrum of disease i.e., Type 1, Type 2, Type 3, Type 4, Type 5. Dengue is spread by variety of female mosquitoes mostly of the Aedes type, mainly A. aegypti. Symptoms include high fever, headache, skin rashes, muscle and joint pain, Nausea, Fatigue, Pain behind eyes. In severe cases there are life threatening symptoms include serious bleeding and shock. Aedes type female mosquitoes when they bite a person infected with dengue during 2 to 10 days gestation period, the virus affects cells lining in gut of mosquito. About 8-10 days later, the virus spreads to other tissues of mosquito affecting salivary glands. The infection of virus does not show any detrimental

effect on mosquito for life. After infection if mosquito bites a healthy person virus enters host's skin through saliva and binds to white blood cells, where the virus reproduces as the blood cells move throughout the body.

Aedes aegypti are particularly involved because they lay eggs in artificial water containers near proximity of humans as they prefer blood meal from humans than from other vertebrates.

Dengue is transmitted from human to human through blood transfusion and organ donation though they have low chance to transmit dengue fever. In Singapore transmission of dengue through blood transfusion is 1.6 to 6 for 100 transfusions. Background analysis shows that not only above conditions are causes for dengue, Climatic factors like rainfall, humidity, temperature, degree of urbanization and quality of vector control services (QVC's) plays key role for dengue outbreak. More rainfall leads to high breeding of mosquitoes and temperature is very important for mosquito lifecycle. Predicting a dengue outbreak is needed for regional governments either to prevent it from happening by using QVC's or by providing surplus medication for outbreak. Statistical methodologies such as gradient descent with bagging and boosting can be used to predict dengue outbreak accurately. Figure 1.1 shows the symptoms of dengue.

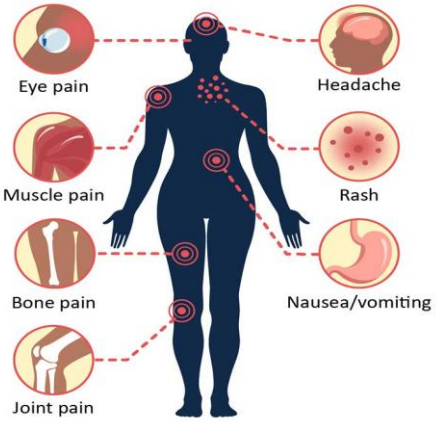

Fig. 1.Symptoms of Dengue (Source: cdc.gov)

Manuscript published on November 30, 2019.
* Correspondence Author
Dr. J. Avanija*, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India. Email: avans75@yahoo.co.in
Dr. G. Sunitha*, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India. Email: gurramsunitha@gmail.com
R. Hitesh Sai Vittal, Department of CSE, Sree Vidyanikethan Engineering College, Tirupati, India. Email: hiteshvittal@gmail.com

II. RELATED WORK

A new machine learning approach used to predict dengue outbreak in and around Kuala Lumpur had been proposed by the authors [1]. Daily mean temperature, daily mean rainfall and enhanced vegetation index (EVI) were used as factors in prediction.

Linear regression model was trained based on these three factors. Extra dengue outbreak contribution factors were not taken into consideration and best combination of predictive factors are not identified. The authors in [2] specified the key attributes used to predict dengue such as severity caused due to infection, risk of infection, the timing of dengue and incidence of outbreaks of dengue in a locality. A model was built to predict dengue outbreak based on high resolution weather data. This model consists of two stages

a) Logistic regression model used to predict occurrence of Aedes aegypti larvae in region of interest
b) Linear regression model used on high resolution weather data is used to predict dengue outbreak.

Climate affects significantly on mosquito population according to this model. Accuracy is less as it consists of two stages. Regression analysis is used to estimate the inter-functional relationship between the variables using simple statistical computation in dataset [3]. Chi-square sampling distribution was used to fit data to model. Regression analysis and Gaussian distribution were used to predict and forecast data. The above regression analysis and forecasting model had been used for seasonal disease forecasting and is less accurate. Autoregressive incident model analysis (ARIMA) is used to predict and forecast a week before dengue outbreak in Subang Jaya, Malaysia [4]. ARIMA was used to predict and forecast dengue incidence using Box-Jenkins approach based on dengue database. Temperature, Rainfall and Relative humidity are used as factors of prediction. It predicts dengue outbreak week before with high accuracy. All dengue outbreak factors are not taken into consideration for this model.

The database collected from online resources for various geographical locations during the decades spanning between 1960's to 2010's, had been analyzed by using Markov Chain Monte Carlo sampling for sampling of data and Boosting regression for statistical analysis of data collected from geographical areas of incidence of dengue outbreak [5].

Dengue incidence in a district can be estimated by using ensemble of dengue prediction models of a district together with its neighboring districts [6]. Dengue incidence model for each district was trained using predictor attributes such as previous dengue incident values and Climatic factors (mean temperature, relative humidity and cumulative rainfall). Accuracy of the model and its error was calculated by using Mean Squared Error (MSE). All the neighboring districts taken for this model must be temporally correlated.

Rainfall, temperature and humidity are identified as the major environmental factors which involves in spread of dengue. Regression model had been proposed by considering the major environmental attributes for predicting the outbreak of dengue. The accuracy of the method was measured and seems to be low since the method considered only minimal number of environmental attributes [7]. Chikungunya and dengue are the vector borne diseases caused to epidemic spread. The analysis of vectors for the outbreak is still under research and the mapping of these vectors based on geographical distribution is required to be planned considering the health of public. Based on the data collected region wise maps were used for prediction of disease based on the impact of the Aedes viruses [8].

A SVR model with cross-validation technique is proposed to develop an effective predictive model for dengue cases [9]. Experimentation is done on the data of 2014 dengue outbreak in china.

III. METHODOLOGY

Data Pre-processing carried out through the phases such as Data Cleaning, Data Editing, Data Reduction and Data Wrangling. Data Cleaning is process where corrupt or inaccurate or incomplete data in a record is replaced with new values either by using mean of the values or by removing the whole record as specified in equation 4.1 where X_i represents the value of an attribute in i^{th} tuple and n represents number of tuples. As data is entered manually or acquired mechanically, missing values in records occur. Data editing is process that involves reviewing, adjustment or updation of collected data from a survey. Selective editing is best method to edit data as experimental or instrumental error otherwise known as outliers. In data either through experimental errors or instrumental errors outliers are generated and can be detected during data editing as specified in equation 4.2. Data reduction is a process of transformation of data, acquired through experimentally or empirically, into a simplified form in a range (generally in range of 0 and 1) which in layman's language called as Normalization. Z- Score normalization method is used in the proposed system as specified in equation 4.3 where x is random member, μ is mean of population and σ is standard deviation. Data wrangling is a transforming raw data and mapping to another data format. This includes munging, data aggregation and data visualization. Data visualization is used for statistical analysis and data aggregation is used to refine data to be trained into a model.

$$Mean = \frac{\sum_{i=0}^{n} Xi}{n} \quad (1)$$

$$Q_1 = \frac{1}{4}(n+1)^{th}\ term \quad (2)$$

$$Q_3 = \frac{3}{4}(n+1)^{th}\ term \quad (3)$$

$$Q_2 = Q_3 - Q_1 \quad (4)$$

$$Z - score = \frac{x-\mu}{\sigma} \quad (5)$$

Bagging (Bootstrap Aggregation) is an ensemble meta algorithm used for improving accuracy of machine learning algorithm used in regression. By using bagging variance is reduced and is used to avoid over fitting. As bagging is ensemble algorithm many independent prediction models are built to predict and combined using some averaging techniques (average, weighting average).

An independent model may predict a value less accurately compared to ensemble of models. Boosting is an ensemble supervised learning algorithm for reducing bias and variance and for accurate prediction of models. A family of algorithms in supervised learning is used to transform set of weak learners into a strong learner. A weak learner is a classifier that is only slightly correlated to true classification. In contrast strong learner is a classifier that is strongly correlated to true classification. A weak learner has slight correlation to true classification compared to set of weak learners.

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (6)$$

Now,

$$\frac{\partial}{\partial \theta} J_\theta = \frac{\partial}{\partial \theta} \frac{1}{2m} \sum_{i=1} (h_\theta(x_i) - y_i)^2 \quad (7)$$

$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i) . \frac{\partial}{\partial \theta_j} (\theta x_i - y_i) \quad (8)$$

$$\frac{\partial}{\partial \theta} J_\theta = \frac{1}{m} \sum_{i=1}^{m} \left((h_\theta(x_i) - y) x_i \right) \quad (9)$$

Therefore,

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^{m} \left((h_\theta(x_i) - y_i) x_i \right) \quad (10)$$

Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function iteratively. The local minimum of a function can be calculated using gradient descent by taking steps proportional to either negative or approximate gradient of the function considering the current position as specified in equation (10) [10]. Gradient descent predicts values more accurately compared to linear regression.

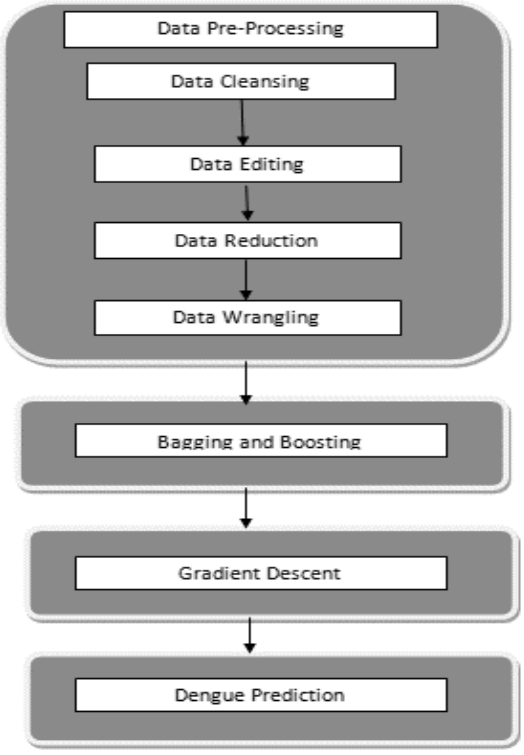

Fig. 2.Flow Diagram for Dengue Prediction Application

A regressor is the one that fits a model using given features and predicts the unknown output value. Dengue dataset from District Health Officer (D.H.O) is collected and four phases of data pre-processing is applied over the collected dataset as specified in figure 2. Figure 3 shows the algorithm for dengue prediction.

Input: Dengue dataset
Output: Probability of Dengue endemic.

1. Input dataset is checked for missing values and calculated mean is replaced in place of missing value.
2. After cleaning Input dataset is checked for outliers.
3. After outlier detection, all the features values are normalized using Z-Score normalization
4. After normalization data is wrangled.
5. Initialize estimation value as 0.0.
6.
   a) Bagging and Boosting: Split dataset into groups and each group is trained using cross validation technique to develop an independent model for each group.
   b) Calculate Mean Squared Error (MSE) with true values to predicted values.
   c) Through error detection classify independent models as weak learners and strong learners.
   d) Depending on classification, weight is assigned to each model.
   e) Weighting average is calculated as

$$Weighted\ Mean = \frac{\sum_{i=0}^{n} (Wi * Xi)}{\sum_{i=0}^{n} (Wi)}$$

7. Calculate modified error for each cluster for every prediction.
8. Repeat the steps from step 6 until accurate prediction result is obtained.

Fig. 3.Algorithm for Dengue Prediction

IV. EXPERIMENTAL RESULTS AND DISCUSSION

Dengue Datasets from 2015 to 2018 are acquired from district health office (D.H.O) of Chittoor district, Andhra Pradesh, India.

Table-I: Attributes of Dengue Dataset

Attributes of Dengue Dataset
Average Temperature
Average Humidity
Average Rainfall
Altitude
Population density
Average Dengue incidence
Vegetation Index

Table I shows the features present in dataset such as Average Temperature, Average Humidity, Average Rainfall, Altitude, Population density, Average dengue incidence, Vegetation index.

The 2001-2011 dengue dataset is divided into three subsets for training, validation and testing. Three different models are applied on datasets with different ratios of training, validation and testing data subsets: one model with the previous average dengue incidence and average temperature as inputs, and the other one with average dengue incidence and average rainfall, and other one with average dengue incidence and average humidity.

Table-II: Training Data Vs Test Error

Training Data (%)	Validation Data (%)	Test Data (%)	Test Error (No of Dengue cases per 21 days) (%)
60	20	20	7.2
70	15	15	6.6
80	10	10	6.2
90	5	5	6.8

After each model is trained, the model is fit into the best regularization parameter to avoid overfitting. Table II shows the mean test errors for these three different models on different runs. Table III shows mean test errors of different clusters.

Table- III: Error based on Clustering of data

Humidity		Temperature		Rainfall	
Group1	Error Rate (%)	Group 2	Group1	Error Rate (%)	Group 2
Cluster 1	6.5	Cluster 1	Cluster 1	6.5	Cluster 1
Cluster 2	5.9	Cluster 2	Cluster 2	5.9	Cluster 2
Cluster 3	7.1	Cluster 3	Cluster 3	7.1	Cluster 3

From Table II, best ratio of training validation and testing dataset is taken and average test error ratio for each cluster is calculated.

Group1 (average dengue incidence, average humidity)
Group2 (average dengue incidence, average temperature)
Group3 (average dengue incidence, average rainfall)

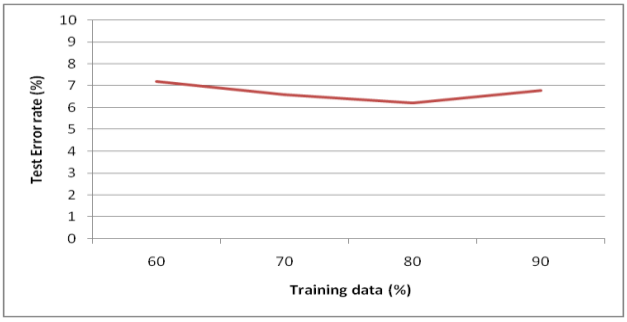

Fig. 4. Training data and Test error rate

V. CONCLUSION

Dengue is a chronic disease caused by the Aedes mosquito in human body which may cause death if proper treatment is not followed. A vaccine termed as Dengvaxia for dengue fever has been approved and is available in most of the countries commercially. Prediction of non-contagious disease such as dengue is a difficult task. This model will predict the number of dengue cases that can occur, based on provided humidity, temperature, vegetation index and rainfall. If there are any chances of occurrence of outbreak then alerts will be sent to nearer health care organizations. Deep learning architectures can be used to enhance the prediction of dengue and decrease test error rate percentage.

REFERENCES

1. Khoo, Jing Jing, Dickson Lukose, Sazaly bin Abu Bakar, and Abdul Sattar, An Efficient Method to Predict Dengue Outbreaks in Kuala Lumpur", 2015.
2. Hettiarachchige, Chathurika, Stefan von Cavallar, Timothy Lynar, Roslyn I. Hickson, and Manoj Gambhir, Risk Prediction System for Dengue Transmission Based on High Resolution Weather Data, PloS one 13(12), 2018.
3. Khadar Babu Shaik, N. Ramachandra and M. V. Ramanaiah, Robust Regression Model for Prediction and Forecasting of Dengue Fever Attacked in Rural Areas of Andhra Pradesh, India, International Journal of Pure and Applied Bioscience, vol.6(1), 2015, pp.318-321.
4. Dom, Nazri Che, A. Abu Hassan, Z. Abd Latif, and Rodziah Ismail, Generating Temporal Model using Climate Variables for the Prediction of Dengue Cases in Subang Jaya, Malaysia, Asian Pacific Journal of Tropical Disease, vol.3 (4), 2013, pp.352-361.
5. Bhatt, Samir, Peter W. Gething, Oliver J. Brady, Jane P. Messina, Andrew W. Farlow, Catherine L. Moyes, John M. Drake et al., The Global Distribution and Burden of Dengue, Nature, vol. 496 (7446), 2013, pp. 504-.
6. Loshini, T., Vijanth S. Asirvadam, Sarat C. Dass, and Balvinder S. Gill, Predicting Localized Dengue Incidences using Ensemble System Identification, In 2015 International Conference on Computer, Control, Informatics and its Applications (IC3INA), IEEE, 2015, pp. 6-11.
7. Lai, Y., The Climatic Factors Affecting Dengue Fever Outbreaks in Southern Taiwan: An Application Of Symbolic Data Analysis, Biomedical Engineering Online,vol.17 (2), 2018, pp. 148-.
8. Kraemer, Moritz UG, Marianne E. Sinka, Kirsten A. Duda, Adrian QN Mylne, Freya M. Shearer, Christopher M. Barker, Chester G. Moore et al., The Global Distribution of the Arbovirus Vectors Aedes Aegypti and Ae. Albopictus, Elife, vol. 4, 2015.
9. Guo, Pi, Tao Liu, Qin Zhang, Li Wang, Jianpeng Xiao, Qingying Zhang, Ganfeng Luo et al., Developing a Dengue Forecast Model using Machine Learning: A Case Study in China, PLoS Neglected Tropical Diseases, vol. 11 (10), 2017.
10. Mathulamuthu, Shermon S., Vijanth S. Asirvadam, Sarat C. Dass, Balvinder S. Gill, and T. Loshini, Predicting Dengue Incidences using Cluster Based Regression on Climate Data, In 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), IEEE, 2016, pp. 245-250.

AUTHORS PROFILE

Dr J. Avanija is Associate Professor ofof CSE at Sree Vidyanikethan Engineering College, Tirupati, India. She received her Ph.D. degree in CSE from Anna University, Chennai. Her area of interests: Semantic Web Mining, Machine Learning, Big Data and Cloud Computing.

Dr. G. Sunitha is professor of CSE at Sree Vidyanikethan Engineering College, Tirupati, A.P., India. She received her Ph.D. degree in CSE from S.V.University, Tirupati. She has 19 years of experience in academia. Her research interests include Data Mining, Spatio-Temporal Analytics, Machine Learning and Artificial Intelligence. She has 7 patents and 3 books published to her credit. She has published around 30 papers in reputed journals and conferences. She has been serving as reviewer for several journals; and has served on the program committees and co-chaired for various international conferences.

R. Hitesh Sai Vittal is a student of B. Tech. CSE at at Sree Vidyanikethan Engineering College, Tirupati, A.P., India. His area of interests include Machine Learning and Internet of Things.