WEB BASED SYSTEM FOR DENGUE EPIDEMICS PREDICTION BASED ON METEOROLOGICAL AND SURVEILLANCE DATA

[1]Snehal Bhate, [2]Shubhangi Khandekar, [3]Atmaja Dhumal, [4]Priya Kumari, [5]Prof.Dr.S.P.Kadam

***Abstract*:** Dengue is one of the major public health problems in India. Early prediction of a Dengue outbreak is the key for control of dengue morbidity, mortality as well as reducing the risk of transmission of dengue in the community and can help policymakers, health providers, medical officers, ministry of health and other health organizations to better target medical resources to areas of greatest need. Here developed model "Web Based System for Dengue Epidemics Prediction Based on Meteorological and Surveillance Data" can help as an early warning tool to identify potential outbreaks of dengue. In this study two popular data mining classification algorithms Random Forest (RF) and K-Nearest Neighbors (KNN) are used for Dengue prediction using a large dataset of Pune city. Data of all 16 wards of Pune city, from 2017 to 2019 has been considered. Parameters used are Average monthly rainfall, Temperature, Humidity, Total number of positive cases and outbreak occurs in binary values (Yes/No). Data samples were collected from Pune Municipal Corporation and India Meteorological Department. Root Mean Square Error (RMSE) and Receiver Operating Characteristic (ROC) are used to measure the performance of the models. It is observed that performance of the model developed using Random Forest is more accurate than KNN. The Random Forest model can predict the outbreak 30-40 days in advance. However accuracy of prediction can be increased using more training data. This model can be scaled-up at country level.

Keywords: Classification, Naïve Bayes, Random Forest, K-Nearest Neighbors, Dengue.

Introduction:

Dengue fever is a mosquito-borne tropical disease caused by the dengue virus. Dengue is spread by several species of female mosquitoes of the *Aedes* type, principally *A. aegypti*. The virus has five types; infection with one type usually gives lifelong immunity to that type, but only short-term immunity to the others. Subsequent infection with a different type increases the risk of severe complications. A number of tests are available to confirm the diagnosis including detecting antibodies to the virus.

Aedes is a genus of mosquitoes originally found in tropical and subtropical zones but now found on all continents except Antarctica. Some species have been spread by human activity others *Aedesalbopictus*, a most invasive species, was recently spread to the New World, including the United States, by the used-tire trade. According to the planet Health Organization, dengue contagion has enlarged 30-crinkle worldwide concluded the previous five eras. Around 50100 gazillion innovative contagions happen every year in additional than 80 nations. Antiviral drugs and vaccines have yet to reach the market and initial results from trials have been discouraging. In the absence of an effective vaccine against dengue, suppressing the dengue cases population remains the key thrust of dengue-control program [1].

Although dengue is a notifiable disease in India, studies and modeling estimates suggest that the disease is grossly under-reported. Another study reported that the actual number of cases in the country were 282 times the number reported by the national vector-borne disease control program [2].

Related work:

An important research objective is to develop models that enable, or enhance, forecasting of outbreaks of dengue, giving medical professionals the opportunity to develop plans for handling the outbreak, well in advance [3]. Due to the exponential growth of available data for dengue forecasting systems, and data storage become important issues as they require potentially hugely greater computational power to handle the greater volume of data, and more complex models

[3]. Agarwal, N., Koti, S. R., Saran, S. and Kumar, A. S. [4] study explains the adopted multi- regression and Naïve Bayes approach to model the relation between dengue cases and weather parameters, i.e. maximum temperature, rainfall and relative humidity [4].

The SVR model had the consistently smallest prediction error rates for tracking the dynamics of dengue and forecasting the outbreaks in other areas in China [5]. Though the SVM classification approach is applied for the prediction analysis. The SVM classifier has less accuracy, as it is not able to drive relationship between the attributes and target set. To increase accuracy of dengue prediction the technique of SVM can be replaced with the voting based classifier [6].

Shakil, K. A., Anis, S. and Alam, M. has done prediction that Naïve Bayes is the best performance algorithm 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 [7].

Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K. and Kesorn, K. [8] paper has been focused four stages namely pre-processing, attribute selection, clustering and predicting the dengue fever. R 3.3.2 Tool is used for pre-processing the household of dengue dataset. 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 [9] [10].

Jain, R., Sontisirikit, S., Iamsirithaworn, S. and Prendinger, H. [11] this study allows for combining different predictors to make forecasts with a lead time of one month and also describe the statistical significance of the variables used to characterize the forecast. This paper focuses more on the socio-economic parameters and how it affects the rate of dengue and it also informs how one should protect themselves from these socio-economic causes of dengue [11].

Motivation:

Dengue is a perilous disease that kills 218 people on average in India throughout a year. Early detection of disease progression associated with severe dengue, and access to proper medical care lowers fatality rates of severe dengue to below 1%. The predictive models are very reliable. Such research into integrating knowledge with Machine Learning is of great interest to many epidemiological and computer scientists and future work should develop models for more accurate predictions.

Theoretical Framework:

DATA SOURCES FOR DENGUE FORECASTING MODELS

In the creation of forecasting models, the dataset is an important element to ensure relevant and usable outcomes. The process of forecasting will necessarily use historical information, and requires a large amount of data, sufficient to support learning to inform the forecasting models. In addition, there must be enough data to allow valid testing of predictions against observed historical events. So, the volume and variety of data is important as is the correctness of the data from reliable sources. Dengue is the most important arboviral disease globally, and the transmission of dengue is closely linked to climate [12].

Typically, and traditionally, data comes from government institutions, such as the Ministry of Public Health, Meteorological Department, Ministries of Agriculture or Lands Administration, and other formal institutions such as hospitals. However, there is no single source that can provide data covering all aspects of forecasting model construction. Now, in the information age, enormous volumes of data are available from new data sources on the Internet, and social. Here, we refer to data sources as being conventional data sources, or modern data sources [13]. For this model, meteorological parameters are considered from India Meteorological Department and the local dengue cases are considered from Pune Municipal Corporation.

DATA CLEANSING

Data mining is defined as the process in which useful information is extracted from the raw data. In order to acquire essential knowledge it is essential to extract large amount of data. This process of extraction is also known as misnomer [6]. In the case of data integration from multiple sources, data is often redundant with different formats, making it necessary to do data consolidation or integration of the data from the various sources by eliminating the redundant and unused data. Data Cleaning/Cleansing or Data Scrubbing is a very important process in predictive modeling. If the analysis of the relationship of variables is done on poor quality data, models derived from the analysis are not reliable. A researcher should identify the data cleansing process used in medical information, including the type and error rate of the data, so that they are fully aware of the implications of the data. The majority of data errors are caused by aggregation of data from multiple sources and integration of this information has several problems, which directly affect the predictive power of the model.

DENGUE FACTORS

Researchers around the world have conducted research on dengue prediction in many countries and various factors have been used to study the correlation between those factors and the number of cases, the factors mostly used are mean temperature, relative humidity and rainfall [5][14]. The objective of their studies is to find the factors that are significantly correlated to the number of cases and deploy those factors in the forecasting model. As a result, the forecasting model can effectively work on predicting the severity of outbreaks in the future.

Therefore, policy makers or responsible staff can be warned in advance and can prepare protocols and all relevant resources for the coming epidemic. Meteorological parameters like, average rainfall, min/max humidity and min/max temperature were collected from the year 2017 to 2019 on a monthly basis from India Meteorological Department and local dengue cases were collected from 16 wards across Pune city with the help of Pune Municipal Corporation.

MACHINE LEARNING TECHNIQUES FOR FORECASTING MODEL CONSTRUCTIONS

The relationship between dengue cases and meteorological features is highly complex and cannot be easily fitted by the classical time series model [15]. Our entire methodology is divided into three approaches: Naïve Bayes approach, random forest and spherical k-nearest neighbors. As the dataset consists of numerical values, the random forest gives the number of cases that can occur given the weather conditions. Based on historical data, the data is discretized into outbreak and non-outbreak using k-nearest neighbors method.

The Naïve Bayes technique is a data mining technique which belongs to the probabilistic classifier. Naïve Bayes is widely used in text classification tasks and it assumes that the attributes are distributed independently. It is based on Bayes' theorem which is

$$p(C_k \mid x) = \frac{[p(c_k)p(x \mid C_k)]}{[p(x)]}$$

Figure 1: Naive Bayes Theorem

where x = (x1, … , xn), C = k possible classes, p(x) = probability of x, $p(C_k)$ = probability of class k, $p(C_k|x)$ = conditional probability of class k for given x, and $p(x|C_k)$ = conditional probability of x for given class k. The dataset is divided into training and test datasets to detect outbreak. Maximum temperature, rainfall, relative humidity and population density are used as attributes. The Bayes technique is chosen due to its efficiency of parameter estimation from small datasets [4].

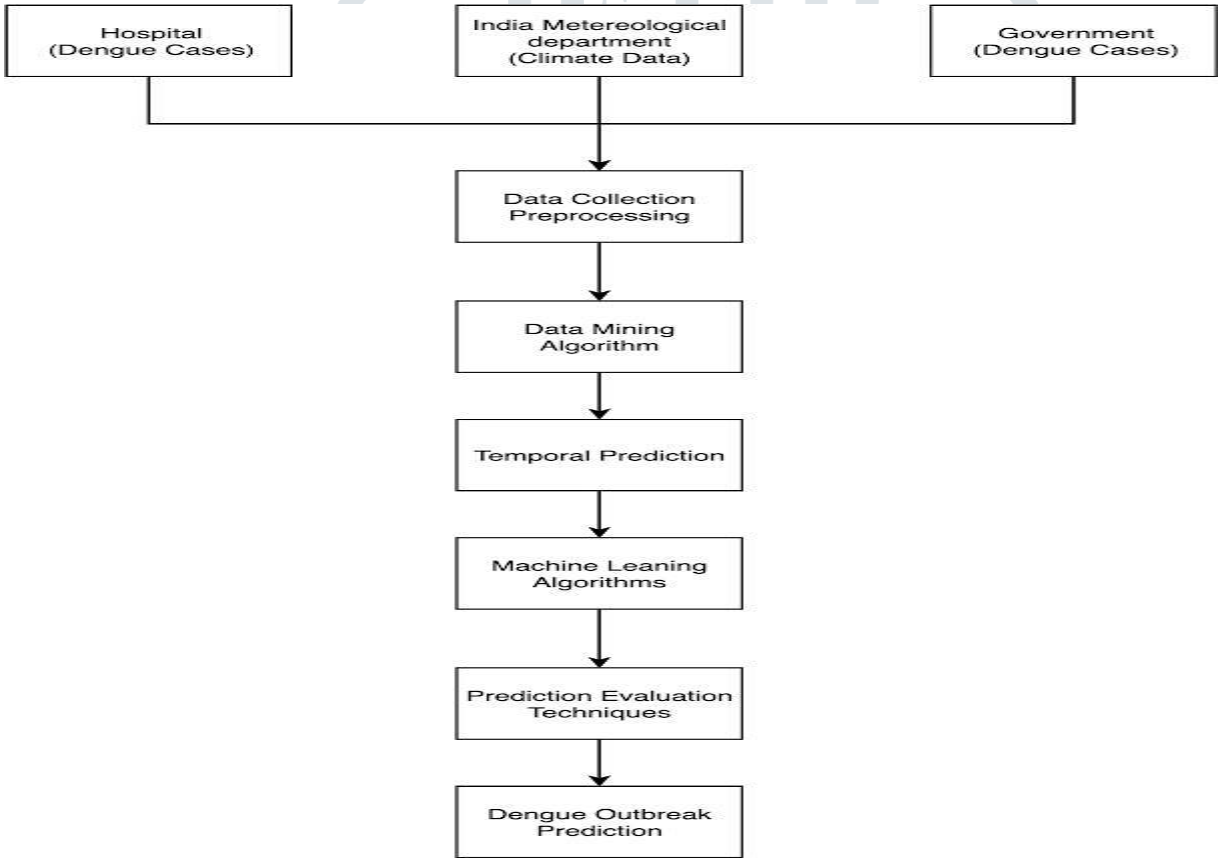

Figure 2: System Architecture

Random Forest algorithm are an ensemble supervised learning method which is used as predictor of data for classification and regression. In the classification process algorithm build a number of decision trees at training time and construct the class that is the mode of the classes output by using each single tree. Random Forest algorithm is a grouping of tree predictors where each tree based on the values of a random vector experimented independently with the equal distribution for all trees in the forest. Presenting the accurate kind of randomness makes them accurate classifiers and regression. Single decision trees often have high variance or high bias.
Random Forests tries to moderate the high variance problems and high bias by averaging to find a natural balance between the two extremes. Considering that Random Forests have few parameters to tune and can be used simply

with default parameter settings, they are a simple tool to use without having a model or to produce a reasonable model fast and efficiently.

Random Forests produces several classifications for given trees. Each tree is grown as follows:
1. If number of circumstances in the training data set is D, sample D cases at random state but with replacement, from the original dataset. This sample testing set will be the training set for increasing the tree.
2. If there are input variables from training dataset, a number is indicated such that at each node of the tree, m variables are selected at random available for the and the best splitting on these is used to splitting the node. The value of is used as constant during entire the forest growing.
3. Each tree is grown to the largest size as possible. There is no pruning an overall grownup tree.
K− NN is another method used to create dengue prediction models. The output from k-NN can be both categorical and numerical data. K-NN classifies data based on the distance among data in the vector space. By using the Nearest Neighbor Index, spatial information of the risk areas of dengue infection have been analyzed using dengue hemorrhagic data from 1998 to 2004, indicating dengue movement patterns from rural communities to urban communities in Trinidad [13].

Conclusion:

According to WHO, early detection of disease progression associated with severe dengue, and access to proper medical care lowers fatality rates of severe dengue to below 1%. The proposed methodology can help as an early warning tool to identify potential outbreaks of dengue almost 30-40 days ago. Predicting dengue outbreaks with in a city on the basis of wards can be done with the help of meteorological parameters and surveillance data, and the results are statistically significant.
In the future, with some more information like the accurate rainfall according to location and population density could be considered and with the help of those parameters we can find descriptive results such as cases according to different types of dengue viruses. This model can also be scaled up at country level in future.

References:

[1] Shi, Y., Liu, X., Kok, S., Rajarethinam, J., Liang, S., Yap, G., Chong, C., Lee, K., Tan, S. S. Y., Chin, C. K. Y., Lo, A., Kong, W., Ng, L. C. and Cook, A. R. 2016. Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore. Environment Health Perspect, 124(9): 1369-1375.

[2] Murhekar, M. V., Kamaraj, P., Kumar, M. S., Khan, S. A., Allam, R. R., Barde, P., Dwibedi, B., Kanungo, S., Mohan, U., Mohanty, S. S., Roy, S., Sagar, V., Savargaonkar, D., Tandale, B. V., Topno, R. K., Sapkal, G., Kumar, C. P. G., Sabarinathan, R., Kumar, V. S., Bitragunta, S., Grover, G. S., Lakshmi, P. V. M., Mishra, C. M., Sadhukhan, P., Sahoo, P. K., Singh, S. K., Yadav, C. P., Bhagat, A., Srivastava, R., Dinesh, E. R., Karunakaran, T., Govindhasamy, C., Rajasekar, T. D., Jeyakumar, A., Suresh, A., Augustine, D., Kumar, P. A., Kumar, R., Dutta, S., Toteja, G. S., Gupta, N. and Mehendale, S. M. 2019. Burden of dengue infection in India, 2017: a cross-sectional population based serosurvey. Lancet Glob Health, 7(8): e1065-e1073.

[3] Stanaway, J. D., Shepard, D. S., Undurraga, E. A., Halasa, Y. A., Coffend, L. E., Brady, O. J., Hay, S. I., Bedi, N., Bensenor, I. M., Castaneda-Orjuela, C. A., Chuang, T., Gibney, K. B., Memish, Z. A., Rafay, A., Ukwaja, K. N., Yonemoto, N. and Murray, C. J. L. 2016. The global burden of dengue: An analysis from the Global Burden of Disease Study 2013. Lancet Infectious Diseases, 16(6): 712-723.

[4] Agarwal, N., Koti, S. R., Saran, S. and Kumar, A. S. 2018. Data mining techniques for predicting dengue outbreak in geospatial domain using weather parameters for New Delhi, India. Current Science, 114(11): 2281-2291.

[5] Guo, P., Liu, T., Zhang, Q., Wang, L., Xiao, J. and Zhang, Q. 2017. Developing a dengue forecast model using machine learning: A case study in China. PLoS Neglected Tropical Diseases, 11(10): e0005973.

[6] Taneja, P. and Gautam, N. 2019. Hybrid Classification Method for Dengue Prediction. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, 8(6).

[7] Shakil, K. A., Anis, S. and Alam, M. 2015. Dengue disease prediction using Weka data mining tool. Proceedings of IIRAJ International Conference (ICCI-SEM-2K17), GIFT, Bhubaneswar, India, ISBN: 978-93-86352-38-5.

[8] Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K. and Kesorn, K. 2016. Analysis of significant factors for dengue fever incidence prediction. BMC Bioinform, 17(166): 1-9.

[9] Sharma, K. D., Mahabir, R. S., Curtin, K. M., Sutherland, J. M., Agard, J. B. and Chadee, D. D. 2014. Exploratory space-time analysis of dengue incidence in Trinidad: A retrospective study using travel hubs as dispersal points,19982004. Parasite Vectors, 7(1): 341.

[10] Manivannan, P. and 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), Srivilliputhur, 1-5.

[11] Jain, R., Sontisirikit, S., Iamsirithaworn, S. and Prendinger, H. 2019. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect Disease. 19(1): 272.

[12] Xu, Z., Bambrick, H., Frentiu, F. D., Devine, G., Yakob, L., Williams, G. and Hui, W. 2020. Projecting the future of dengue under climate change scenarios: Progress, uncertainties and research needs. PLoS Neglected Tropical Diseases, 14(3): e0008118.

[13] Siriyasatien, P., Chadsuthi, K., Jampachaisri, K. and Kesorn, K. 2018. Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes. IEEE Access, 6: 53757-53795.

[14] Jaafar, Ibnu, Abidin, Z., Norhaslinda, Jamil M. and Jastini. 2016. Modelling the prediction of dengue outbreak using system dynamics approach. Jurnal Teknologi, 78: 107-113.

[15] Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L. and Liu, Q. 2020. Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. Int J Environ Res Public Health, 17(2): 453.