International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206544

Feature extraction and prediction of Dengue Outbreaks

Kunal Parikh[1], Tanvi Makadia[2], Harshil Patel[3]
Information Technology Department, A. D. Patel Institute of Technology, Karamsad, Gujarat, India

Article Info
Volume 6, Issue 5
Page Number: 216-222
Publication Issue :
September-October-2020

Article History
Accepted : 01 Oct 2020
Published : 14 Oct 2020

ABSTRACT

Dengue is unquestionably one of the biggest health concerns in India and for many other developing countries. Unfortunately, many people have lost their lives because of it. Every year, approximately 390 million dengue infections occur around the world among which 500,000 people are seriously infected and 25,000 people have died annually. Many factors could cause dengue such as temperature, humidity, precipitation, inadequate public health, and many others. In this paper, we are proposing a method to perform predictive analytics on dengue's dataset using KNN: a machine-learning algorithm. This analysis would help in the prediction of future cases and we could save the lives of many.

Keywords: Machine Learning, K-Nearest Neighbors, Prediction, Dengue, Meteorological data

I. INTRODUCTION

Dengue is a viral fever transmitted by female mosquitoes. We currently have no vaccine or specific course of treatment for dengue. Dengue outbreaks have been rapidly increasing, the number of dengue cases reported to WHO increased over 8 fold over the last two decades, from 505,430 cases in 2000 to over 2.4 million in 2010, and 4.2 million in 2019. To decrease the impact of such outbreaks and be equipped to deal with it effectively we can use modern technology. Researches indicate a strong relation between dengue outbreaks and meteorological data. Through this paper, we want to throw insight into how we can use Machine learning (KNN model) to predict the potential dengue outbreak so that it can be contained at the rudimentary stage.

The dengue virus (DEN) belongs to the family Flaviviridae and genus Flavivirus. It consists of four distinct serotypes (DEN-1, DEN-2, DEN-3 and DEN-4) which are closely related to each other. The Aedes aegypti mosquito is the primary vector that sends the infections that cause dengue. The infections are transmitted to people through the bite of an infective female Aedes mosquito, which predominantly passes the infection while feeding on the blood of an infected individual.

Earlier, dengue was limited to few southern states of India, however, from 2001 the total number of dengue cases has significantly increased and has been widely spread across India. Initially, dengue was spread in urban areas only, but now it has proliferated to rural areas as well. The expansion of dengue is mainly because of environmental changes, immunological factors of community, rapid unplanned urbanization

and host-pathogen interactions. Precipitation and temperature changes are two important climatic factors in the transmission of the disease. Precipitation gives the water that serves as a habitat for larvae and pupae and temperature impact the reproduction of mosquitoes.

According to NASA earth observatory, the temperature has surged 0.6°C to 0.9 °C between 1906 and 2000, and since the past 5 decades, the rate of temperature has increased two times. Moreover, the variation of precipitation in the atmosphere has increased. Warm temperatures and high humidity are favourable for mosquitoes which increases their life expectancy and decreases their incubation period; thus, they replicate faster. Precipitation also plays an important role in the breeding of mosquitoes; it results in a sudden increase of dengue infection during the rainy season. On the contrary, it is observed that dry conditions can also spread infections in urban areas because people with little access to water tend to store water in unprotected areas near their households. Also, various studies have reported changing spatial patterns in the transmission of dengue. This is mainly due to the growth of trade and tourism industries which favours the transmission of the virus.

For predicting the outbreak we will be using Machine learning. Machine learning is a type of artificial intelligence that uses a set of tools for making predictions and getting insights from the data. Machine learning is classified into three categories: supervised learning, unsupervised learning and reinforcement learning. Here in this paper, the algorithm used is a type of supervised machine learning algorithm.

In supervised learning, as the name suggests, there is the presence of a supervisor as a mentor. Here, we train our machine by using well-labelled data. During the training phase, the system is fed with massive amounts of data. The machine is specifically told what to look for and the model is trained until we get our desired accuracy. Once the training is completed on labelled data, we provide the machine with a completely new set of data (validation data) and the machine has to predict the correct outcomes based on its previous learning.

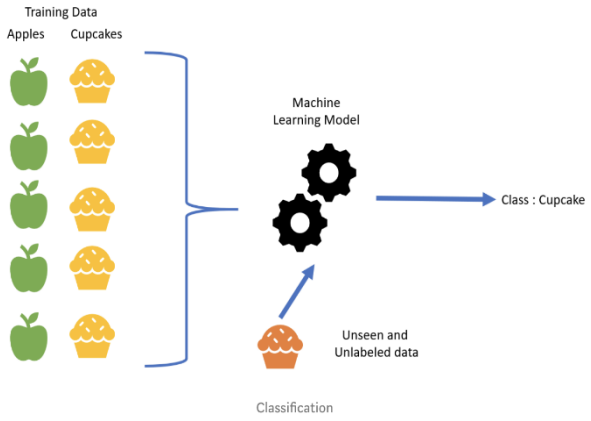

Figure 1.1. Example of Supervised machine learning algorithm

Supervised learning is mainly used for classification and regression problems. A classification algorithm helps us to categorize our data into different categories. For example, determining what category a book belongs to or the category of an object based on its shape. On the other hand, regression algorithms take a different approach. Here, the algorithm expects the model to produce numerical outcomes. For instance, predicting the amount of sales of a store for a future given date or estimating the amount a customer will pay for a certain product.

The K-Nearest Neighbour (KNN) algorithm is a supervised learning algorithm. It depends on a case learning approach that is affected by the lazy learning strategy. KNN is also known as Instance based learning. Instance based strategy is known as memory-based learning. KNN is based on the theory that similar data points lie near each other. In this methodology, it matches a new instance problem

with previous instances that are provided during the training. It gets stored in the memory. When predicting a new data point or classifying a novel data point we pay attention to the K number of near neighbours, because they would be similar to the new data point. Then we can calculate the distance between the novel data point and the individual neighbours. We can then use this distance to calculate the class of the novel data point or the regression value. There are different methods that we can use like mean, mode, median and weighted mean. It is generally productive for colossal datasets that have fewer attributes and provide a universal approximation. Training a KNN model also requires comparatively less time.

The KNN strategy can be applied to both regression and classification. For both the applications, the input consists of k nearest training instances in feature space. In KNN classification, the outcome is a class to which the novel instance belongs to. The classification of an element is chosen as the premise of a dominant vote of its neighbours. Conversely KNN regression, the outcome is the merit significance for the object. The significance is the means of the values of their KNN. In this paper, we will be using the KNN regression model.

To predict the value of a new point using the KNN regression algorithm, the following method should be followed. The first step is to calculate the distance between the new point and each point of the training set. For a continuous variable, we can calculate three types of distance: Euclidean, Manhattan, and Hamming distance. Hamming distance is used for discrete values whereas Euclidean and Manhattan are used for continuous values.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Figure 1.2. Formula to calculate euclidean distance and manhattan distance

$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

Figure 1.3. Formula to calculate hamming distance

Once the distance between the new point and each point of the training set has been calculated, the next process is to pick the closest points. K is defined by the number of neighbouring points to consider while determining the value of a new point.

In the second step, the value of K is to be selected. We can select any number of K we wish, but our goal is to select such a value of K that will predict the most optimum value of the new point. This can be done by looping over finite value of K and selecting the value that gives the least error and maximum accuracy score. We can locate the ideal estimation of K by utilizing the k-overlap cross-approval. It includes assessing the test mistake rate by holding out a subset of the preparation set from the fitting cycle.

In order to define the quality of the regression model and predicted values , r^2 score and error values such as mean absolute error and root mean square error are used to estimate the difference between the true values and the predicted values.

The absolute error is the difference between the predicted values and the actual values. Thus, the mean absolute error is the average of the absolute error.

$$\text{Mean Absolute Error} = \frac{1}{N}\sum_{i=0}^{n} |y_i - \hat{y}_i|$$

Next is the root mean squared error or RMSE is similar to the MAE, except you take the average of the squared differences between the predicted values and the actual values and roots them.

Because the differences are squared, larger errors are weighted more highly, and so this should be used when you want to minimize large errors. Below is the equation for MSE and rooting it we get RMSE

$$\text{Mean Squared Error} = \frac{1}{N}\sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

Further, R Squared(r^2) is a measurement that tells you to what extent the proportion of variance in the dependent variable is explained by the variance in the independent variables. In simpler terms, while the coefficients estimate trends, R-squared represents the scatter around the line of best fit.

$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation}$$

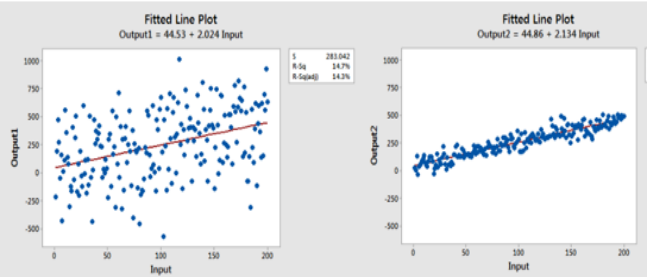

Figure 1.4. Low r^2 score vs. high r^2 score

For example, if the r^2 is 0.70, then 70% of the variation can be explained by the model's inputs.

If the r^2 is 1.0 or 100%, that means that all movements of the dependent variable can be entirely explained by the movements of the independent variables.

The error values must be as low as possible and the r^2 score must me near to one for best predicted results

KNN algorithm for datasets that have a lot of features or have most features whose value is zero is not very efficient. The KNN is sensitive to noise in the data because it calculates the mean to k nearest neighbours. We need to eliminate the less significant features because the difference between the neighbours can be escalated by unnecessary features.

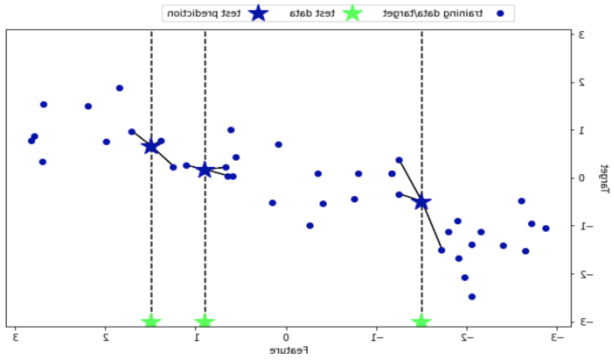

Figure 1.5. KNN Visualisation

II. WORKFLOW

A. Data Gathering and Visualization

First, the data was gathered. The datasets were taken from Kaggle. One contained the cases of dengue of a city in India over the span of 10 years in a weekly format and the other consisted of the features to be used for prediction like temperature, humidity etc. reported in different weeks and over the years in the city. Libraries such as NumPy, pandas, Matplotlib and seaborn were used to plot the visualization of the cases. The dataset in CSV format was read and converted into data frames using the pandas library. Then the data was plotted in the form of a line graph using seaborn. The cases were grouped by years and aggregated to obtain the total cases in a particular year. The following graph illustrates the trend of dengue cases over the years. The timeline was plotted on the X-axis and a number of cases were plotted on the Y-axis.

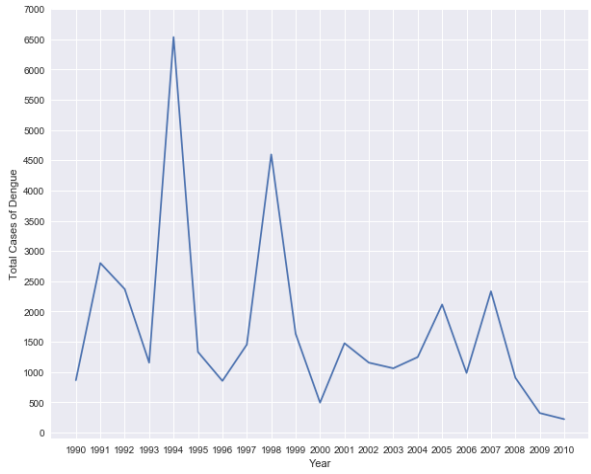

Figure 2.1. Plotting Dengue Cases

B. Data Pre-processing and Cleaning

Srinivasa Rao Mutheneni, Andrew P Morse, Cyril Caminade, Suryanaryana Murty Upadhyayula in their paper "Dengue burden in India: recent trends and importance of climatic parameters'' have clarified that overall humidity/precipitation and temperature are some of the key factors for the dengue outburst in a particular year. The next step involved processing and cleaning of the data. Data cleaning is a process in which one goes through all of the data within a dataset and either remove or update information that is incomplete, incorrect, improperly formatted, duplicated, or irrelevant. This helps in removing ambiguity and improving results. The dataset of the features had some inaccurate and missing values. The columns having the missing values were first identified and listed and then the values were filled using the mean method. The dataset consisted of the precipitation amount, maximum air temperature, minimum air temperature, maximum station temperature and minimum station temperature over the weeks of the years. The temperature was converted from Kelvin to Centigrade and all the values of different columns were rounded off up to 3 decimal places to maintain uniformity. The maximum station temperature in Celsius and minimum station temperature in Celsius of different weeks of a particular year were combined as average station temperature and similarly the maximum air temperature and minimum air temperature were combined as average air. Furthermore, the total cases of columns were added to the data frame.

1990	0.1226	0.103725
1990	0.1699	0.142175
1990	0.03225	######
1990	0.128633	0.245067
1990	0.1962	0.2622
1990	######	0.17485
1990	0.1129	######
1990	0.0725	0.0725
1990	######	0.146175

Figure 2.2. Data before cleaning

1990	0.1226	0.103725
1990	0.1699	0.1317
1990	0.03225	0.142175
1990	0.128633	0.245067
1990	0.1962	0.2622
1990	0.18245	0.17485
1990	0.1129	0.0882
1990	0.0725	0.0725

Figure 2.3. Data after cleaning

C. Training and Feature Engineering

The next step involved the feature engineering part-process of using domain knowledge to extract features from raw data via data mining techniques. These features were used to improve the performance of machine learning algorithms. Dummy variable traps i.e. a scenario in which two or more variables are highly correlated were removed. This step was followed by splitting the data frames into training sets and testing sets (x_train, x_test, y_train and y_test). The test size used was 0.2. In addition to this feature, scaling was done to normalize the range of independent variables or features of data by using the StandardScaler of sklearn model.

D. Modelling, Value of K and Result

The final step involved the modelling process with KNN by using the KNeighborsRegressor module from the sklearn neighbours model.

```
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from math import sqrt
from sklearn.neighbors import KNeighborsRegressor

mae_val = [] #to store Mean Absoulute error values for different k
rmse_val = [] #to store root mean square error values for different k
r2_val = [] #to r^2 values for different k
for K in range(10):
    K = K+1
    model = KNeighborsRegressor(n_neighbors = K)

    model.fit(x_train, y_train)  #fit the model
    pred=model.predict(x_test) #make prediction on test set
    error = mean_absolute_error(y_test,pred)#calculate mae
    r2error = r2_score(y_test,pred)#calculate r^2 score
    rmse = sqrt(mean_squared_error(y_test,pred))#calculate rmse
    mae_val.append(error) #store mae values
    r2_val.append(r2error) #store r^2 values
    rmse_val.append(rmse)
```

Figure 2.4. Implementation

The n_neighbors was looped from the range of 0 to 10. For different value of K, the r^2 score, root mean square error and absolute mean error was first stored in an array and then plotted to obtain the best value of k where the error rate was minimum and the r^2 score was maximum. Following were the results obtained for different value of K.

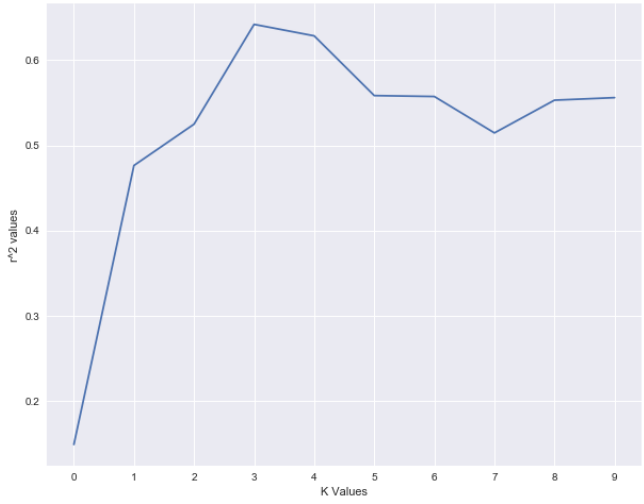

Figure 2.5. r^2 score for different values of K

The best r^2 value was obtained at the k value of 3 with the score of 0.6154

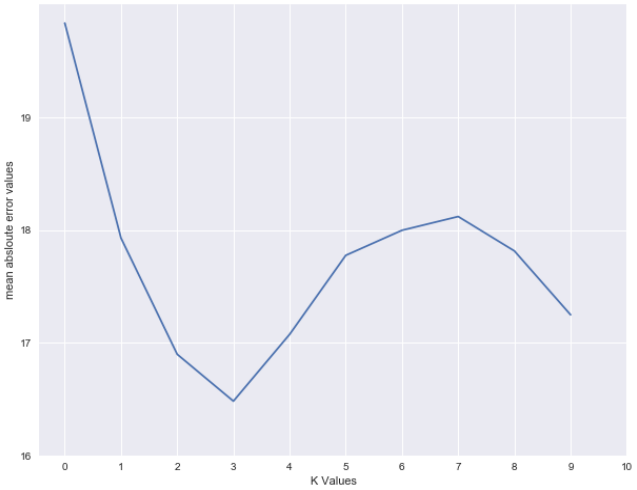

Figure 2.6. Mean absolute error for different values of K

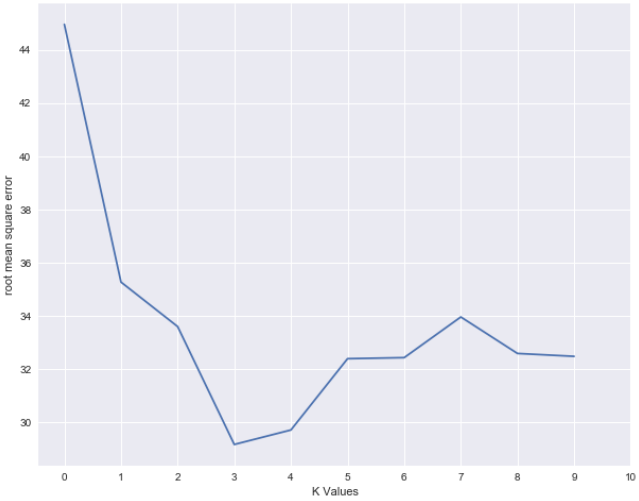

Figure 2.7. Root mean square error for different values of K

The lowest absolute mean error and root mean square error was observed at k= 3 with error of 16.637 and 29.145 respectively.

III. CONCLUSION

Dengue is a fatal viral disease. Successful prediction of dengue outbreak can lead to early precautions to prevent the disease and reduce the number of deaths. It has been established that dengue outbreaks are affected by the meteorological factors of the area. In this paper we have been able to design a dengue prediction model using KNN. The factors considered are temperate and humidity over the period .Our

model was able to predict the cases accurately and the results obtained were quite good. We were able to get r^2 score of 61.54%, the mean absolute error of 16.637 and root mean square error of 29.145 for the k value of 3.

IV. REFERENCES

[1] Srinivasa Rao Mutheneni,Andrew P Morse,Cyril Caminade,Suryanaryana Murty Upadhyayula, "*Dengue burden in India: recent trends and importance of climatic parameters*"

[2] Aditya Lia Ramadona,Lutfan Lazuardi,Yien Ling Hii,Åsa Holmner,Hari Kusnanto,Joacim Rocklöv, "*Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data"*

[3] Sadegh Bafandeh Imandoust, Mohammad Bolandraftar, "*Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background* "

[4] Jason Brownlee, *"K-Nearest Neighbors for Machine Learning*"

[5] Raghvendra Jain, Sra Sontisirikit, Helmut Prenginger, "*Prediction of dengue outbreaks based on disease surveillance, meteorological and socioeconomic data*"

[6] S. Appavu, MohamedMallick, G.Chinthana, "*Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking"*

[7] Wei-Chun Tseng, Chi-Chung Chen, Ching-Cheng Chang, Yu-Hsien Chu "*Estimating the economic impacts of climate change on infectious diseases: a case study on dengue fever in Taiwan*"

[8] P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, K. Kesorn "*Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes*"

[9] J E T Akinsola "*Supervised Machine Learning Algorithms: Classification and Comparison*"

[10] Ayush Kumar Rathor, Dr.RanjanaRajnish "*Comprehensive Review of Data Visualization Techniques using Python*"

Cite this article as :

Kunal Parikh, Tanvi Makadia, Harshil Patel, "Feature extraction and prediction of Dengue Outbreaks", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 216-222, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT206544
Journal URL : http://ijsrcseit.com/CSEIT206544