International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

The Generic Framework for Prediction of Epidemic Disease Outbreak

Priya Bansal[1], Mohammed Asif[2], Smitha G.R.[3], Sharadadevi S Kaganurmath[4]

[1]*Department of Information Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*
[2]*Department of Information Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*
[3]*Assistant Professor, Department of Information Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*
[4]*Assistant Professor, Department of Information Science and Engineering, R.V. College of Engineering, Bengaluru, Karnataka, India*

Abstract - *This paper analyses the correlation of disease outbreak with climate conditions of the regions. The modelling of number of cases reported historically in a region using the climatical conditions resulted in the increase in accuracy in order to predict the spread of disease outbreak. This would prove useful for the government in proactive disease management and also to plan prevention and treatment measures.*

The dengue outbreak in a city is predicted based on the factors such as precipitation, temperature, humidity and other climatic factors. The paper deals with the forecast of dengue based on various models such as Simple Linear Regression model, Negative Binomial regression and Arima model. The construction of the model also includes feature selection and feature engineering approaches in order to obtain improved results.

Key Words*: *Linear Regression; Negative Binomial Regression; Auto Regressive Integrated Moving Average(ARIMA)

1.INTRODUCTION

In recent times the infectious diseases present a constantly changing threat to public health. Epidemics cause significant social, economic, and health impact on societies. The impact of infectious diseases on the human population is a function of many factors that includes transmission mechanics, vector dynamics, environmental conditions, transmission mechanics, social and cultural practices. The geography of the place such as latitude and the diversity of flora and fauna in the region, the interaction between the species, the habitat of the animals has the major contribution in the origin of the infectious disease. The population of the region, transmission rate and the immunity of the people towards the disease decides whether the infectious disease would take the form of pandemic. It is noticed that many of the deadly infectious diseases are mostly concentrated in certain regions of the world. This depends mainly on the climatic conditions of the regions, i.e certain diseases outbreak in rainy regions when there is enough moisture to transmit easily.

1.1 CHALLENGES FACED IN DISEASE OUTBREAK PREDICTION

There are multiple models to forecast the seasonal as well as emerging disease outbreak. But the choice to select the reliable model to emulate the disease behaviour remains a major challenge. The Real time prediction of the number of dengue cases may help researchers make a reliable prediction even if people did not report or go to the hospitals. These models can help to reliably forecast the future path of the outbreak can provide additional insight to develop and control the epidemic by timely actions and necessary changes in their policies in order to effectively manage the outbreak of the disease. However, creating predictions in real-time poses computational, logistical and statistical challenges. The raw data needs to be available in a standard format for processing into analytical dataset in order to overcome logistical challenges [4].

Thus, the real time information on dengue outbreak in the cities of United States was taken into consideration which is maintained by the Centers for Disease Control and Prevention.

1.2 PROPOSED DESIGN APPROACH

The most critical phases in proposed model is that of design. In this phase, model architecture and flow is visualised and designed in a way to make it compatible with both the functional and non-functional requirements. Some of the design challenges were assumptions, dependencies and design constraints. These constraints are explained in the later section.

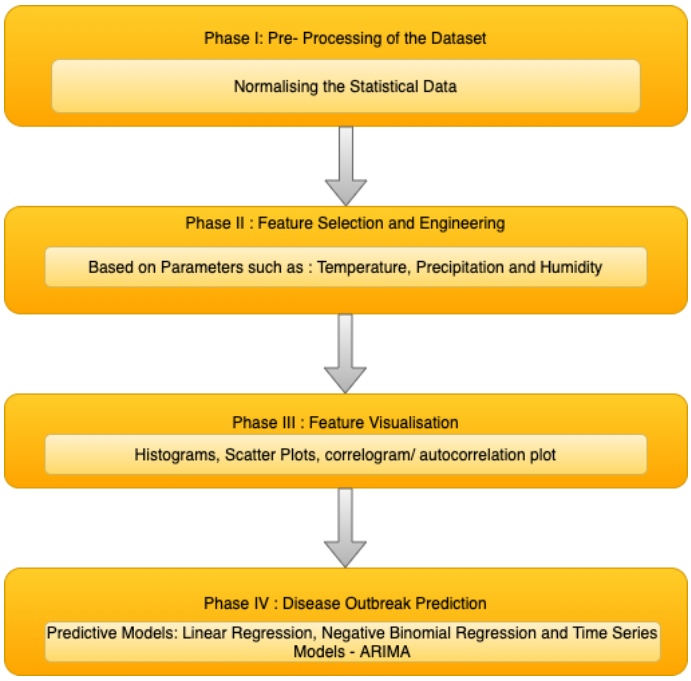

Figure 1.1: Proposed design describing the flow of the-model

2. METHODOLOGY AND DESIGN CONSTRAINTS

The design described in stages are as follows:

Stage 1: The DengAI dataset obtained from drivenData.org was used for initial exploration where all the missing values were replaced with mean values of the column in order to remove any kind of noises present in the dataset. The parameter includes temperature, precipitation, humidity and their average analysis [1].

Stage 2: A different approach to get a perspective on the data was carried out using feature engineering and feature section. The month was extracted out from the 'week start date'. Few of the other feature engineering approaches includes adding city feature as a boolean value, time shifting the data by few steps as there were gaps between the change in climate, mosquito bites and the reporting of the disease. Some improvements were also achieved through feature selection approaches. The results of these approaches have been discussed later in the report.

Stage 3: These feature visualisation then provided us prominent results that helped in analysing the correlation of disease outbreak with the change in climate. The visualisation plots included histograms, scatter plots and correlogram or autocorrelation plots for time-series model [3].

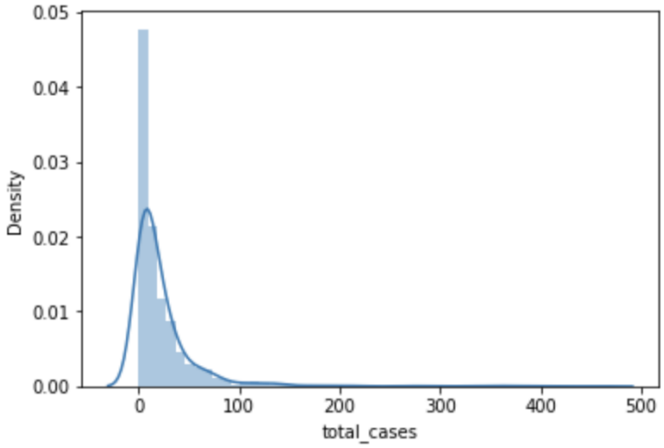

Figure 2.1: Initial data exploration with seaboard displot

The above plot suggests that the total number of cases per week have a median value of 25. It was also observed that the distribution was highly skewed. Further, as the approach suggested to model the number of cases on the basis of climate change, the annual pattern of the disease should suggest a correlation between climate and the total number of cases.

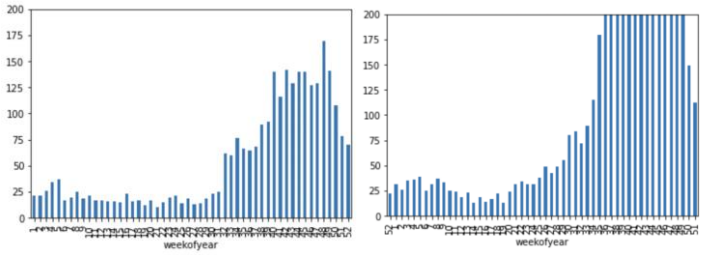

Figure 2.2: Examining years patterns for different cities

Thus, the above plot indicated the possibility of some correlation between the climate and spread of the dengue disease.

Stage 4: The predictive model used for the disease outbreak prediction includes:

(i) Simple Linear Regression: This was used as a baseline model. Regression is a technique for predicting a target value using independent predictors. This method is commonly used for forecasting and determining cause and effect relationships among variables. At first, when the random model was used for computation, it was predicted that the number of cases for any given week was equal to the mean number of cases in the past and also, the MAE was found to be low. However, when Recursive Feature Elimination was implemented by creating a base classifier to evaluate a subset of attributes. The MAE significantly dropped lower. Eliminating

some unnecessary features later proved to improve the accuracy, and applying some regularisation also helped in addressing the overfitting problem.

(ii) Negative Binomial Regression: Negative binomial regression is implemented using maximum likelihood estimation. The dependent variable in a negative binomial regression is a count of the number of times an event happens. For over-dispersed count results, where the conditional variance exceeds the conditional mean, negative binomial regression can be used. Since it has the same mean structure as Poisson regression and an additional parameter to model over-dispersion, it can be called a generalisation of Poisson regression. The confidence intervals for Negative binomial regression are likely to be smaller than those for a Poisson regression model if the conditional distribution of the outcome variable is over-dispersed. It was observed after previous plots, that the data was highly skewed. As the count data was negatively skewed, negative binomial regression outperformed previous discussed models. Since a time lag was found between climate change, mosquito bites, and diseases getting reported, these time lags were filled by introducing data shifting by few more steps. All preprocessing functions have been updated accordingly to allow seamless testing of this concept. However, this distribution of the cases didn't give the desired results.

(iii) ARIMA Model: ARIMA models are a type of statistical model that was used to analyse and forecast time series results. The following are the parameters used for the ARIMA model:

- p: The lag order, or the number of lag observations used in the model.
- d: The degree of differencing is the number of times the raw observations are differenced.
- q: The order of moving average, also known as the height of the moving average window.

The aim of this model, which is based on an adjustment of observed values, was to minimise the variance between the values provided in the model and the observed values as close to zero as possible. This model was used to explain the action of both stationary and non-stationary sequences, giving situational variances more flexibility. The results of these models have been discussed later in a separate section.

3. RESULTS AND DISCUSSIONS

As a baseline model, Simple linear regression and its variants are used from their implementation in scikit learn along with regularisations. It was observed that the data is highly skewed. Also, when the count data is negatively skewed, it is observed that negative binomial regression outperforms other models. However, ARIMA model's fixed structure provides an edge when the data is generated by a process similar to ARIMA assumptions.

The results of these models have been discussed below, in detail.

(i) Simple Linear Regression:

The Mean Absolute Error (MAE) is the loss function that is used in a regression model. It calculates the average degree of errors in a series of forecasts without taking into account the directions of the errors.The range is also 0 to ∞.

METHOD USED	MAE	RESULT OBTAINED
Feature Extraction : Extracted Month from date	18.2	The resulted graph was unable to show even a slight linear relationship between the actual and predicted values.
Feature Selection	9.6	Recursive feature elimination helped to improve the accuracy.

Figure 3.1: Tabular Representation of the Results.

At first, when the random model was used for computation, we predicted the number of cases for any given week to be equal to the mean number of cases in the past. The mean absolute error came to be 18.2. However, when simple linear regression was used for the implementation the MAE dropped to 12.4 which signified that it was better than random model. Recursive Feature Elimination was then implemented by creating a base classifier to evaluate a subset of attributes. The MAE significantly dropped to 9.6. Eliminating some unnecessary features helped us to improve the accuracy. Also, applying some regularisation helped to address the overfitting problem.

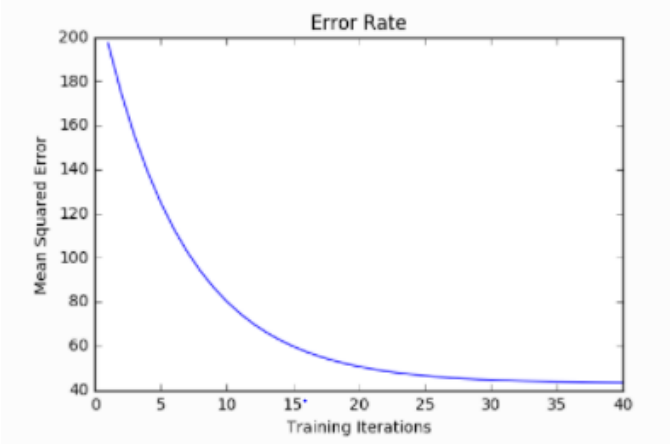

Fig. 3.2 Relationship between number of iterations and mean square error

However, a simple linear model is only slightly better than a random model as the scatter plot suggests that this model isn't good as there isn't even a slightly linear relationship between the actual and predicted values.

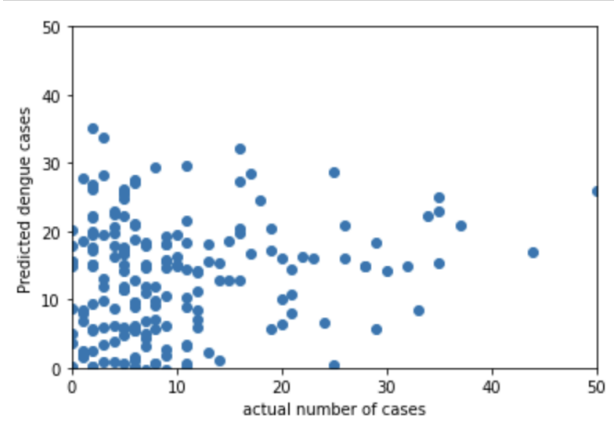

Fig. 3.3 Relationship between actual and predicted number of cases

(ii) Negative Binomial Regression: It was observed after previous plots, that the data was highly skewed. As the count data was negatively skewed, negative binomial regression outperformed previous discussed models. Since a time lag was found between climate change, mosquito bites, and diseases getting reported, these time lags were filled by introducing data shifting by few more steps. All preprocessing functions have been updated accordingly to allow seamless testing of this concept. However, this distribution of the cases didn't give the desired results.

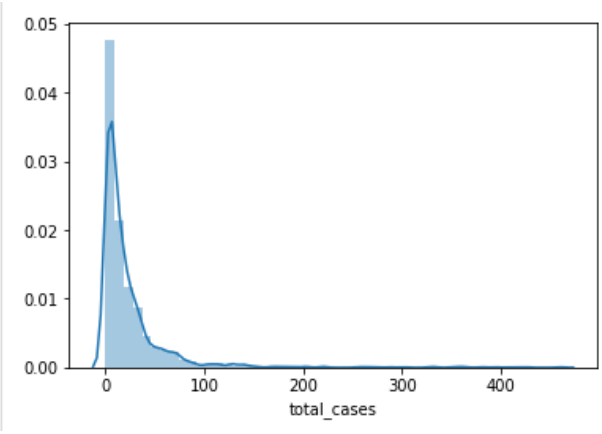

Fig. 3.4 The distribution of dengue cases

It can be seen in the graph, that y labels have a highly skewed distribution, which, when combined with the assumption that the labels are count variables, provides a strong argument for using negative binomial regression. The conclusion that is drawn after implementation of the model is that the negative binomial is by far the best model we have come across and also we made the observation that time shifting data actually decreases the performance of the model.

(iii) ARIMA Model: The aim of this model, was to explain the action of both stationary and non-stationary sequences, giving situational variances more flexibility. When the mean and variance of a series remain constant over time and the value of covariance depends only on the distance between two time intervals, the series was said to be stationary.

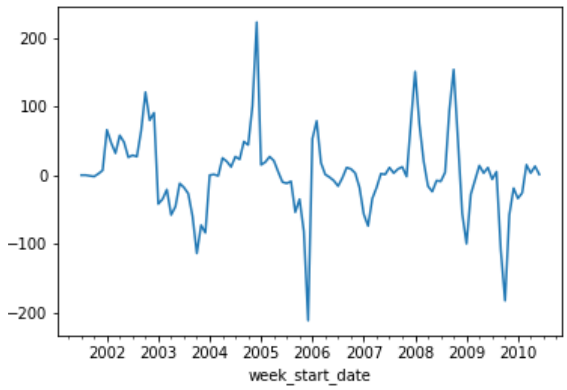

Fig. 3.5 Observing the seasonal differences

This figure shows the optimised results after fitting the model fitting the model on hyper parameters. After computation and comparison, the mean absolute value was obtained to be around 60. The result however, can still be

improved with additional data engineering which can be the future scope of improvement.

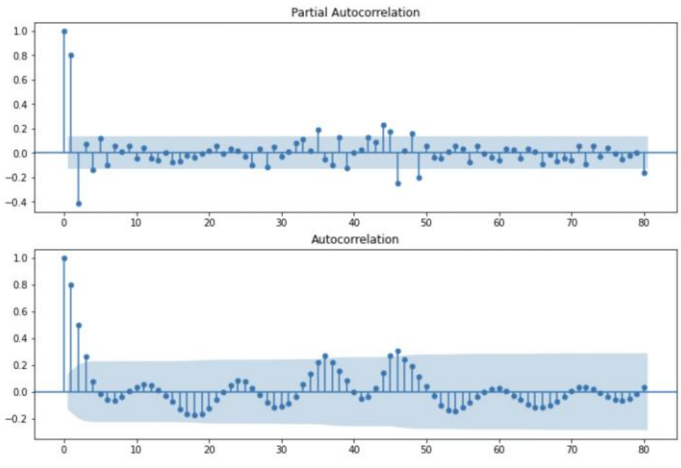

Fig. 3.6 Observing the autocorrelation with lags

This states that periodicity occurs at every year i.e; every 12th month and there is a significant autocorrelation at lags upto 2 or even 4. Also, there is a partial auto correlation until lag 3. This gives a significant amount of correlation of disease outbreak with climate conditions of the region. This would further be helpful in disease outbreak prediction in a specified region.

4. CONCLUSIONS

In this paper, there has been a close observation towards the time series dataset for disease outbreak prediction. This paper analyses the correlation of disease outbreak with climate conditions of the regions. The dengue outbreak in a city is predicted based on the factors such as precipitation, temperature, humidity and other climatic factors. This paper deals with the forecast of dengue based on Simple Linear Regression model, Negative binomial regression and ARIMA model. The model construction also includes feature selection and feature engineering approaches in order to obtain improved results. This study of dengue outbreak prediction derives results that lead to these conclusions :

i. Feature engineering approach proved to be successful when the month was taken in consideration instead of 'week start date' and it performed better. Also, using city feature as a boolean value proved to give better results.

ii. The gaps between change in climate, mosquito bites and reporting of the disease was when compensated with data shifting, instead of giving the promising results, it worsened the performance.

iii. The Negative binomial regression model outperformed all the models as the data was highly negatively skewed.

iv. The ARIMA model stated that periodicity occurred every year i.e; every 12th month. This gave a significant amount of correlation of disease outbreak with climate conditions of the region. The result however, can still be improved with additional data engineering which can be the future scope of improvement.

5. FUTURE ENHANCEMENTS

1. There can be a possibility of improving the scores with more data. However, the computational, logistical and statistical challenges while creating predictions in real time still persists. The raw data needs to be available in a standard format for processing into analytical dataset in order to overcome logistical challenges. Thus, collection and aggregation of data from multiple sources might be a challenge though, it might lead to a route to build this project with more accuracy.

2. Further, the analysis of data from different sources can be accounted even to get the classified results i.e; predicting the infected individuals and classifying them based on their gender in order to understand different aspects in which a specific disease behaves on an individual's body. This would also help to study the mortality rates and accordingly building the vaccination strategies.

3. The use of deep learning might prove to be successful in future cases as recurrent neural networks have been outperforming the traditional algorithms in most of the cases where the data is sequential.

REFERENCES

[1] Aniruddha Adiga, Devdatt Dubhashi, Bryan Lewis, Madhav Marathe, Srinivasan Venkatramanan and anil Vullikanti, " Mathematical Models for COVID-19 Pandemic: A Comparative Analysis", J.Indian Inst. Sci. |VOL 100:4|793-807 October 2020.

[2] Farzaneh S. Tabataba, Bryan Lewis, Milad Hosseinipour, Foroogh S. Tabataba, Srinivasan Venkatramanan, Jiangzhhuo Chen, Dave Higdon, and Madhav Marathe, "Epidemic Forecasting Framework Combining Agent-Based Models and Smart Beam Particle Filtering", IEEE International Conference on Data Mining.

[3] Sasikiran Kandula, Jeffrey Shaman, " Near-term forecasts of influenza-like illness- An evaluation of autoregressive time series approaches, Epidemics 27 (2019) 41-51.

[4] R. Sanjudevi, D. Savitha, " Dengue Fever Prediction using Classification Techniques", Volume: 06 Issue: 02| Feb 2019.

[5] N. A. M. Molinari, I. R. Ortega-Sanchez, M. L. Messonnier, W. W. Thompson, P. M. Wortley, E. Weintraub, and C. B. Bridges, "The annual impact of seasonal influenza in the us: Measuring disease burden and costs," Vaccine, vol. 25, no. 27, pp. 5086–5096, 2017.

[6] Haines L, Munoz W, Van Gelderen C. ARIMA modelling of birth data. J Appl Statistics. 1989;16(1):55-67.

[7] Vanishree K, Nagaraja G.S, Emerging Line of Research Approach in Precision Agriculture: An Insight study,IJACSA publication.

[8]Dupont, W. D. (2019). Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. New York: Cambridge Press.

[9] M. L. Zwilling, "Negative Binomial Regression," The Mathematica Journal, 2018. dx.doi.org/10.3888/tmj.15-6. K. Elissa, "Title of paper if known," unpublished.

[10] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipstick,"Real-time influenza forecasts during the 2017-2018 season."Nature communications, vol. 4, p. 2837.

[11] P. Dawson, R. Gailis, and A. Meehan, "Detecting disease outbreaks using a combined bayesian network and particle filter approach," Journal of Theoretical Biology, vol. 370, pp. 171|183.

[12] F. S. Tabataba, P. Chakraborty, N. Ramakrishnan, S. Venkatramanan, J. Chen, B. Lewis, and M. Marathe, "A framework for evaluating epidemic forecasts," BMC Infectious Diseases, Vol. 17, no. 1, p.345, 2017.

[13] J. Straub, T. Traylor, Nicholas Snell, Gurmeet. " Classifying Fake News Articles Using Natural Language Processing to Identify In- article Attribution as a Supervised Learning Estimator", 2019 IEEE 13th International Conference on Semantic Computing (ICSC), 2019.

[14] Bhavika Bhutai, Priyanshu Sehgal, Archana Purwar. "News Dectection using Sentiment Analysis", 2019 12th IC3, 2019.

[15] M. Negnevisky Artificial Intelligence, " A Guide to Intelligence", England: Pearson Education Limited, 2018.

[16] M. Piao. "Discovery of Significant Classification Rules from Incrementally Inducted Decision Tree Ensemble for Diagnosis of the Diseases", Lecture Notes in Computer Science, 2019.

[17] Smitha G R, Darshan R, "Security Issues in Cloud Computing and Risk Assessment", International Journal of Computational Engineering Research (IJCER), ISSN (e):2250 –3005 June 2014, pp 5-10

[18] Ashwin R, Rakesh S S, and Dr. Preethi N Patil, "Sputum Based Pneumonia Detection through Image Processing", National Conference on Security, Privacy and Analytics –SPA, RVCE, 4th -5th May, 2018.