iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VII **Month of publication:** July 2020

DOI: https://doi.org/10.22214/ijraset.2020.30562

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

Outbreak Prediction of Dengue and Hepatitis A

Miron T Manuel[1], Princy P[2], Riya Reji[3], Alpha Mathew[4]
[1, 2, 3]*Department of Information Technology, College of Engineering Kidangoor, Pala, Kerala, India*
[4]*Asst. Prof, Department of Information Technology, College of Engineering Kidangoor, Pala, Kerala, India*

Abstract: ***An outbreak happens when there is a sudden increase in illness at unexpected high numbers. If we could predict the outbreaks earlier we can take proper planning to avoid the outbreaks. We developed a model 'OUTBREAK PREDICTION OF DENGUE AND HEPATITIS A USING MACHINE LEARNING' which can be put to use as an early warning tool. In this study we use two data mining classification algorithms Support Vector Machine (SVM) and Artificial Neural Network (ANN) are used to forecast or predict an outbreak. The used datasets contains parameters like average estimates of rainfall, respective Temperature, atmospheric moisture or Humidity, Total amount of positive cases and outbreak occur in binary values Yes or No. In this paper, we have not only sought to establish a relationship between climatic components and a possible Outbreak of dengue and hepatitis A, but we also tried to find out which algorithm can be ideal for developing discovered relationships. Root Mean Square Error (RMSE) and Receiver Operating Characteristic (ROC) values are used to measure the models performance. SVM models came out to be more precise than ANN respectively; the precision of the prediction model can be incremented using more training data.***

Keywords: Dengue, Hepatitis A, Support Vector Machine

I. INTRODUCTION

An outbreak is a sudden unexpected increase in the number of cases of a disease. An outbreak used to occur within a region or country, or even may affect multiple countries. It can or may last for a few days or weeks, or months, or even for several years.
Outbreaks of communicable diseases happen with the influence of environmental factors to some extent. Water, hygiene,, food and air quality attributes are key constituents in the transmission and escalation of contagious diseases.
Dengue and Hepatitis A epidemics are occurring repeatedly. Dengue is caused by an Dengue Virus (DENV) which is transmitted by Aedes mosquito. The mosquito life cycle is influenced by various climatic components like temperature, downfall of rain, and relative moisture(humidity). DENV is a single-stranded RNA virus from the Flaviviridae family. It is a zoonotic vector-borne disease transmitted by mosquitoes of the genus Aede. Hepatitis A is an enveloped positive single-stranded RNA virus from the Picornaviridae family of viruses. This virus gets transmitted by contaminated food and water.

II. METHODOLOGY

A. Data Collection

We obtained monthly reported dengue cases ,hepatitis A case data for the years 2011–2019 were obtained from the Health officer and from the kerala DHS website. Also obtained monthly climate data for the years 2011–2019 from the weather forecasting site. We then combined above two data to form our dataset. Average precipitation rate , maximum temperature, minimum temperature, average humidity, amount of positive cases and that of death cases were used as independent variables, and the outbreak reported by the month was taken as the dependent variable.

B. Data Preprocessing

Data preprocessing is the data mining practice that converts raw or unprocessed data into a comprehensible format. Obtained data from the real world, maybe imperfect, insufficient, and differs certain behaviors, and may include many errors. Data preprocessing ensures primary(raw) data is ready for further refining.
Data goes through a few stages during preprocessing:

1) *Data Cleaning:* Missing values are replaced , the noisy data is removed, inconsistencies in the data are taken care of..
2) *Data Integratrity:* Data with different forms are combined together and conflicts within the data are hence resolved.
3) *Data Transformation:* Normalization is performed and then data is generalized.
4) *Data Reduction:* Representation of the data in a data warehouse is being decremented here.
5) *Data Discretization:* Decrementing the number of estimates of a constant trait by splitting the span of attribute intervals.

The Data obtained from different sources were combined together to form the two dataset. One for hepatitis A and other one for Dengue. The dataset is then preprocessed using weka (Waikato Environment for Knowledge Analysis) which being an open source software providing tools for data preprocessing, developing different Machine Learning algorithms , and visualization tools can be of good help in this scenario.

a) *Missing Data:* A few fields contain missing values which might affect the efficiency of the prediction model. For that matter, the missing input was filled using the mean values .

b) *Feature Selection:* In this process we select the best subset of attributes in our dataset. For this we use Pearson's Correlation Coefficient in weka. Table 3.1 contains features selected And Table 3.2 contains samples of data after data preprocessing.

Table 2.1 Input and output features

Input features (X)	1) Monthly Average Precipitation 2)Average Humidity 3)Max Temperature 4)Min Temperature 5)Positive Case 6)Month 7)Death 8)Average Temperature
Output feature (y)	Outbreak (prediction)

Table 2.2 Sample of preprocessed dataset

MONTH	AVG HUMIDITY	AVG TEMP	MAX TEMP	AVG PERCIPITATION	MIN TEMP	POSITIVE CASE	DEATH	OUTBREAK
1	67	27	32	14.6	20	96	0	YES
2	70	28	32	16.6	23	86	1	YES
3	72	29	35	36.1	27	127	1	YES
4	75	29	37	110.9	4	119	1	YES
5	75	28	35	252.6	25	137	1	NO
6	82	27	33	653.2	24	150	0	YES
7	84	26	32	687.2	23	145	1	NO
8	87	26	31	404.7	22	157	1	YES
9	84	28	32	252.3	23	149	0	NO
10	82	28	32	270.7	22	145	0	YES
11	77	28	33	158.6	22	157	1	YES
12	72	28	33	45.9	23	155	2	YES
1	75	27	32	14.6	23	52	0	NO
2	73	28	33	16.6	21	50	0	YES
3	73	29	33	36.1	22	186	0	NO
4	75	29	33	110.9	22	132	0	YES
5	80	28	33	252.6	23	408	1	YES
6	88	27	32	653.2	22	95	1	YES
7	89	26	31	687.2	22	77	0	YES
8	87	26	32	404.7	23	42	0	NO

III. MODEL BUILDING

For building our predictive model we use python.. In python we use pandas for loading dataset and numerical calculations. We obtained machine learning algorithms with the scikit-learn library. Scikit-learn has a frontier that helps to configure all of the power of machine learning excluding troubles.

The machine learning algorithms that perform the job are support vector machine (SVM) and artificial neural network(ANN).

A. Support Vector Machine

Support-vector machines (SVM) is a supervised learning model. SVM contains learning algorithms that examine data used for classification and regression analysis. In case of a set of training examples, each example is marked as affiliated to one of the two categories, an SVM subroutine builds up a model that tabulates new examples to one of the two classes.

The SVM method defines the instances as points in space. Therefore these examples can be separated into different classes with a clear gap. New examples can be then charted into that same space and forecasted to belong to a categorised on the side of the gap on which they fall. Inclusive of executing linear categorization, SVMs can also execute non-linear classifying with the kernel trick, that is increasing the proportional feature spaces of inputs.

SVM constructs a hyperplane in multidimensional space to separate the different classes. A hyperplane is a decision plane that divides a range of objects possessing different class memberships. The basic proposition of SVM is to discover a maximum marginal hyperplane that better splits up the dataset into classes.

Support vectors are the data points that are adjacent to the hyperplane. These points will define the separating line the best by calculating margins. These points are more relevant in the construction of the classifier.

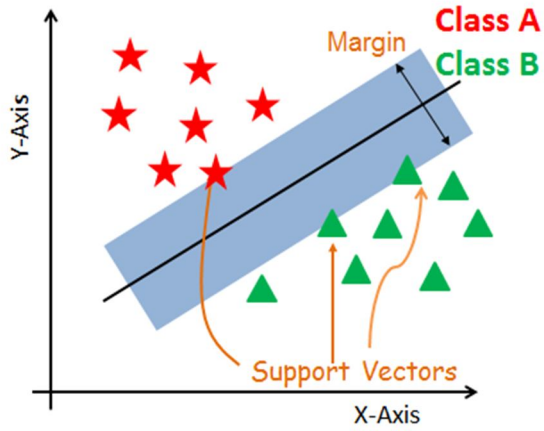


Figure 3.1 SVM Model

Here we are executing the SVM algorithm using a kernel. Kernel converts our input dataset space into the needed form. SVM uses a technique called the kernel trick. Kernel trick ensures in building a more precise classifier.

Linear Kernel A linear kernel can be utilised as a normal dot product of any two specified observations. Product between two vectors is the sum of the multiplication of each set of input values.

$$f(x) = B(0) + sum(ai * (x,xi))$$

B. Artificial Neural Network

It's an supervised learning algorithm. ANNs are inclusive of multiple nodes that replicate biological neurons of the human brain. The neurons are associated via links and the nodes communicate with each other. The nodes extract the input data and exercise operations on it.. The outcome of these operations is designated to other neurons. The output at every node is known as its activation or node value. In the starting, the ANN makes some random forecasting, these are then compared with the correct output and the inaccuracies are estimated.

A neural network implemented in two phases: Feed Forward phase as in figure3.3 and Back Propagation phase.

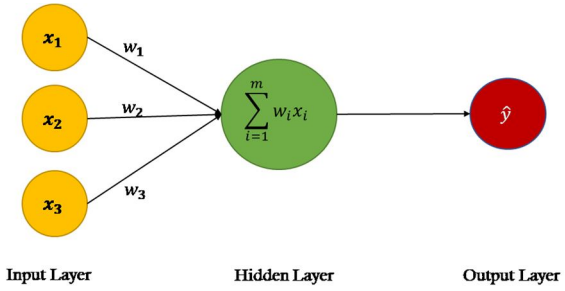


Figure 3.2 ANN's Feed forward phase

Each link has a weight correlated with it. Artificial Neural Network is proficient of learning, which is done by changing weight values. If the network gives rise to a good output, there is no need to modify the weights value. However, if the network gives rise to a poor output or an error, then the system will change the weights value in order to improve successive results.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VII July 2020- Available at www.ijraset.com

IV. CHOOSING THE BEST PREDICTOR

Root Mean Squared Error (RMSE) and ROC (Receiver Operating Characteristic) performance parameters of SVM and ANN model considered for the comparison of accuracy Model build by SVM gave RMSE 0.545 and ROC area 0.881, while ANN produced Root Mean Squared Error 0.63 and ROC area 0.618 on Dengue Data. In the Hepatitis A dataset RMSE value is 0.569 and ROC 0.73 for svm model ,while ANN model developed showed RMSE as 0.699 and ROC area as 0.253. The performance lead for classifications accuracy, says that ROC > 0.80 is observed as a "GOOD" classifier and ROC 0.70 as "FAIR". The Classifier must attain ROC value close to one for higher accuracy of making predictions. For our test dataset the SVM model is making predictions with higher accuracy than the ANN model.

V. RESULT AND CONCLUSION

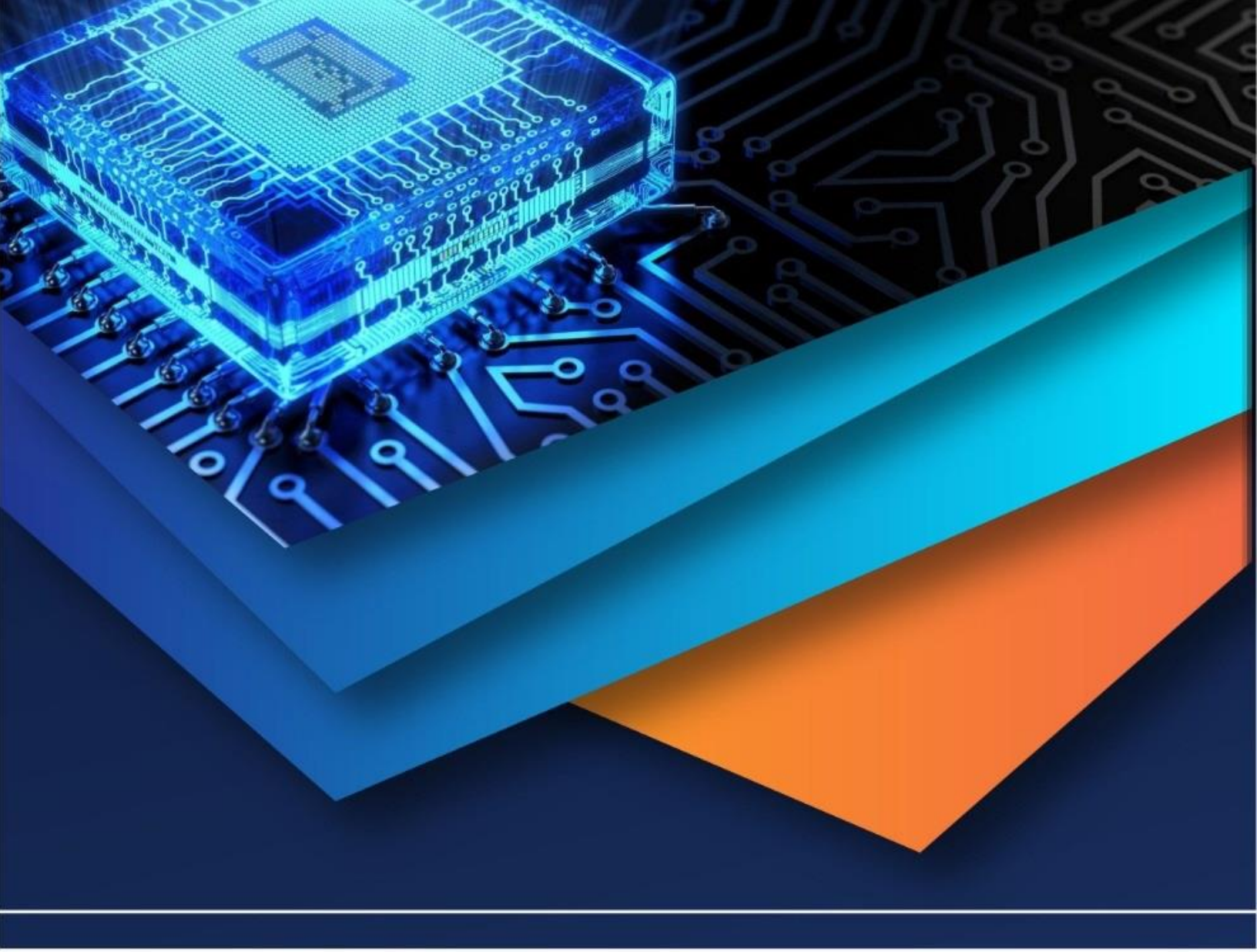
Results implies that the climatic components had a major effect on the happening of dengue outbreak. The model developed could forecast a possible outbreak in advance with considerable accuracy, and this can act as an early warning tool for taking dengue control measures.

The purpose of our project was to predict dengue outbreak cases and Hepatitis A outbreak cases by using the measured real parameters such as monthly average Precipitation values (mm), mean temperature values (°C), mean relative humidity values (percentage) and the amount of dengue and Hepatitis A cases . The performance of the prediction made with the svm classifier was compared with ANN algorithm's prediction results. SVM showed better performance than ANN.

The prediction model has quite constraints in forecasting the outbreak with great accuracy due to lesser data in the dataset used. If more data is used for training, a better prediction model could be developed as increase in training data increases accuracy of prediction. The results obtained shows that a relation between climate changes and occurrence of communicable disease outbreak does exist. So the presented prediction model would be useful for the stakeholders to use as an early warning system in order to control the disease.

REFERENCE

[1] Albert-Schlangen;"Seasonality of Hepatitis: A Review Update",J Family Med Prim Care. 2015 Jan-Mar; 4(1): 96–100.
[2] Rachel Sippy, Diego Herrera, David Gaus, Ronald E. Gangnon, Jonathan A. Patz, Jorge E. Osorio;"Seasonal patterns of dengue fever in rural Ecuador: 2009-2016". May 6, 2019 https://doi.org/10.1371/journal.pntd.0007360 .
[3] Vijeta Sharma, Ajai Kumar, Lakshmi Panat, Dr. Ganesh Karajkhede, Anuradha lele ,"Malaria Outbreak Prediction Model Using Machine Learning",International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 12, December 2015.
[4] P.H.M.Nishanthi Herath, A.A.I. Perera, H.P.Wijekoon; "Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks", International Journal of Computer Applications (0975 – 8887)Volume 101– No.15, September 2014
[5] http://www.cs.waikato.ac.nz/ml/weka/
[6] https://en.wikipedia.org/wiki/Support_vector_machine
[7] https://en.wikipedia.org/wiki/Artificial_neural_network



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY