Prediction of Dengue Fever Outbreaks using Machine Learning Methods

Ponnada Akhil
Department of CSE
GVP College of Engineering(A)

A. Ajaya Kumar
Department of CSE
GVP College of Engineering(A)

ABSTRACT
Mosquitoes are the major source of the spread of dengue. The blood sample of a person is mostly used for detection of dengue. But there are various other factors which are responsible for dengue prevalence.In this project,weather conditions such as dew point, humidity, minimum and maximum temperatures along with precipitation of places present in India are considered to predict whether dengue exists or not.

The four supervised algorithms- k-nearest neighbors, random forest, decision tree and support vector machines are compared to predictions. The results of these algorithms are compared based on accuracy, precision, and recall.

General Terms
Python, Machine Learning, Supervised Algorithms, Dengue.

Keywords
Dengue Prevalence, Machine Learning, SVM, Random Forest, Decision Tree, K-NN

1. INTRODUCTION
Dengue is a mosquito-borne virus transmitted by mosquitoes. Dengue fever is a systemic and dynamic disease that is difficult to predict. Dengue fever is usually similar to other mosquito-borne diseases such as malaria, chikungunya, and leptospirosis[1].Dengue virus mostly spreads at a temperature between 23-28C. In recent statistics from WHO, the global incidence of Dengue has a drastic increase in the last two decades with an estimated infection rate of 100-400 million per year.

Across India, 70 people were killed and more than 36,000 were injured. According to the Ministry of Health, people had been suffering from dengue fever since January. Most infections had been recorded in West Bengal and Orissa in the east, Kerala and Karnataka in the south[3]. An evaluation of warning signs and other relevant medical history is helpful for an early and accurate diagnosis of dengue, especially if you have recently been in a dengue hotspot[2].

In this system,k-nearest neighbor, decision tree, random forest and support vector machines are the machine learning algorithms used for the prediction of dengue based on weather factors such as humidity, dew point, precipitation, minimum and maximum temperatures of places present in India. The best algorithm among these with highest accuracy, precision and recall is used for classification.

2. BACKGROUND AND RELATED WORK
R. Brahmanambika,2018 used Naïve Bayes, Sequential Minimal Optimization(SMO), Random Forest, Reduce Error Pruning(REP) in Weka tool for comparison. In this they got Random Forest as the best algorithm based on accuracy.[3]

M. Mufli,2018 used K-Means Clustering and Support Vector Machine (SVM) Algorithm for classification. In this they got K-Means as the best algorithm based on accuracy.[4]

K.Chellapan,2016 used Decision tree, Discrimination Analysis, SVM, K-NN, an Ensemble classifier for classification on blood profile. In this, they got Discrimination Analysis as the best algorithm based on accuracy.[5].

A. Anitha,2018 used decision tree algorithm for the prediction and got 86.12 accuracy.[6]. Amornsak used Time series analysis and SVM for the prediction of a season in which dengue prevalence is high.[7].

Surya Sumpeno,2020 used Naïve Bayes, Logistic Regression, Fuzzy Logic for prediction. In this they got Fuzzy Logic as the best algorithm based on accuracy.[8]. Husin, proposed a dengue generation prediction architecture using a neural network and a nonlinear regression model that considers only dengue cases and rainfall data.[9].

Dr. G P Saradhi Varma,2018 is credited for using decision trees as a data mining method and proposing a set of useful qualities based on temporal data. The experiment is broken down into four sections. Dengue was classified using a decision tree technique from two independent patient datasets with an accuracy of 97.6% and 96.6 percent [10].

Padet Siriyasatien,1998 suggested a multivariate Poisson regression approach based on statistics. Statistics is a well-established scientific discipline that may be used to validate linear correlations between parameters, whereas data mining methods can be used to uncover hidden knowledge in data[11].

Tsuey-Hwa Hu,1995 suggested a method for predicting dengue risk in real time for a limited region. For early warning, focused monitoring, and action, they employed risk prediction models rather than a typical statistical model. The geographical and temporal units' accuracy may be simply modified to suit various conditions for different cities[12].

2.1 MACHINE LEARNING MODELS FOR ANALYSIS

2.1.1 SUPPORT VECTOR MACHINE

Support Vector Machine is a supervised machine learning algorithm which takes the complete dataset as input and plots each parameter value in an n-dimensional space. Then it generates hyperplanes using thumb rule and picks the best hyperplane which classifies the data. Then it states the class of the data. [13].

2.1.2 RANDOM FOREST

It is a supervised machine learning algorithm which takes the complete dataset as input and divides it into various subsets randomly. It builds decision trees from these subsets. The prediction of each decision tree is considered and the class with highest prediction is given for the test data.[12].

2.1.3 DECISION TREE

This algorithm contains decision nodes and leaf nodes. The decision nodes represent the decision rules with multiple branches and leaf nodes represent the outcome that doesn't have any further branches.

This model takes the whole dataset into the root node and divides it into subsets by extracting the key features using the Attribute Selection Measure(ASM). These subsets again generate decision trees repetitively, until there is no further scope for classification. It stores the predicted classes in the leaf nodes.[11].

2.1.4 K-NN

It is also known as a lazy learning algorithm because it does not immediately learn from the training set, but instead saves the data set and performs an action on the data set at the time of predicting the major class. Here, K is the parameter related to the number of closest neighbors to include in the majority of the voting process.[14].

2.1.5 PRINCIPAL COMPONENT ANALYSIS

PCA uses a linear transformation to reduce the size of the feature set. A new dataset can have: Features equal to or less than the original data set. The covariance matrix is used to calculate the principal components. These components are listed in descending order of importance.

3. RESEARCH METHOD

It describes the used dataset for the detection of dengue prevalence, proposed method to perform analytics, and discusses the evaluation metrics applied on the classification algorithms.

3.1 DATASET

The new data set has total 1457 records in that 75% is given as training data set and 25% is given as testing data set input to the algorithms. The algorithms classify the data and detect the dengue prevalence.

3.2 DESIGN OVERVIEW

Figure 1. describes the proposed model for detection of dengue prevalence that consists of Pre-processing, Classification and Evaluation phases which are explained below,

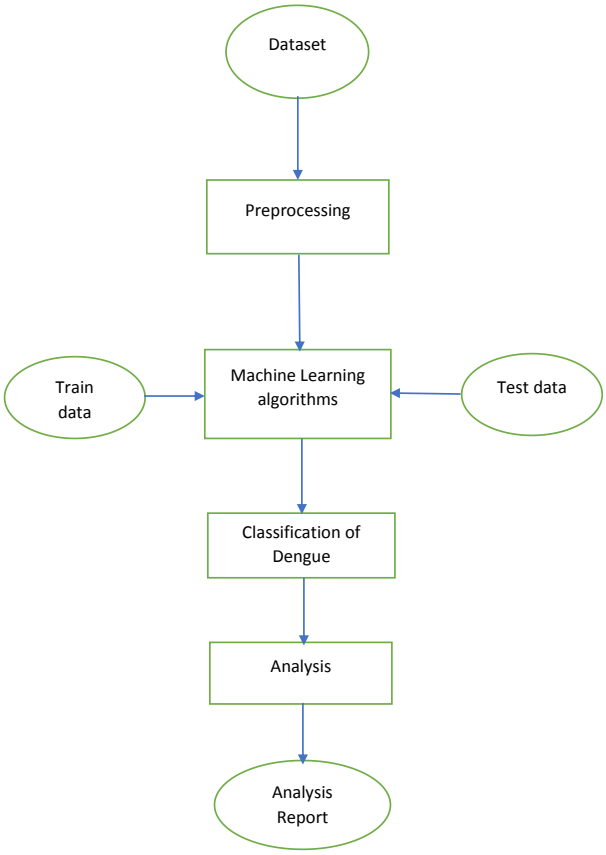

Fig 1: System Architecture

3.2.1 PREPROCESSING

The pre-processing step is necessary for the detection of dengue prevalence. In the proposed model, the missing values present in the dataset are filled using mean values in this step. This results in a dataset with a complete weather condition values that can be given as an input for the classification algorithm.

3.2.2 CLASSIFICATION TECHNIQUES

In this project, the classifier models constructed are Random forest, support vector machine, k-nearest neighbour, and Decision Tree to classify whether dengue prevalence in a particular weather condition or not. These classifiers have been discussed in the previous section, and the evaluation of their performance is carried out in the next section.

4. EXPERIMENTAL RESULTS

In this section, the results of the experiment and their significance are discussed based on the following table, where table1, refers to the complete evaluation metrics results (accuracy, precision, recall, F1-Score) on the four algorithms.

Table 1. Comparison Of Various Machine Learning Algorithms With Respect To The Given Problem Statement.

No.	Algorithm	Accuracy	Precision	Recall	F1 Score
1	Random Forest Classifier	0.9974	0.997	0.997	0.997
2	Decision tree	0.9888	0.983	0.985	0.987
3	K Nearest Neighbors Classifier	0.9338	0.890	0.930	0.909
4	Linear SVC	0.7500	0.730	0.485	0.577

4.1 EVALUATION METRICS

The effectiveness of a proposed model can be determined by applying a few evaluation metrics to calculate how accurately the model can differentiate dengue prevalence or not. In this research, four machine learning algorithms have been constructed, namely Random forest, support vector machine, k-nearest neighbor, Decision Tree. So, to review these models the standard evaluation metrics used by the research community are applied on them.

ACCURACY

Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of model performance, he following equation can be used [8]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Where TP- True Positive
TN-True Negative
FP- False Positive
FN-False Negative

PRECISION

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It is the correct prediction of dengue prevalence out of all predictions[13].

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Where TP- True Positive
FP- False Positive

RECALL

Recallis defined as the proportion of the relevant cases that were actually found among all the relevant cases. It is the correct prediction of dengue prevalence out of all correct predictions [13].

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Where TP- True Positive
FN- False Negative

F1-SCORE

The F1-Score is defined as a harmonic mean of precision and recall[13].

$$F1\ Score = \frac{2*(Precision\ *recall)}{Precision\ +recall} \quad (5)$$

4.2 COMPARISON OF ALGORITHM RESULTS

The proposed model uses four machine learning techniques that were set to achieve better accuracy. From the below graph, it can be inferred that the Random Forest classifier gave the best accuracy, precision, recall and f1-score on thedataset, where accuracy is 97%, precision, recall and f1-score of 97%. While SVC gave same precision but had the slightest differences in their accuracy of 0.96. Logistic Regression gave third best accuracy, precision, recall and F1-Score values while, Naïve Bayes gave least accuracy, precision, recall and F1-Score values among all the four algorithms.

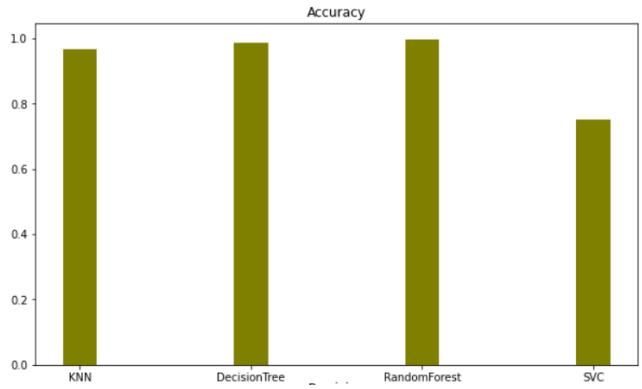

Fig 2: Bar chart of accuracy with respect to four classifiers

From the above figure 2, it can be inferred that, random forest classifier achieved better accuracy of 0.997 and Decision tree gave second highest of 0.988. While K-NN gave third highest

of 0.933 and SVM gave least accuracy value of 0.75 respectively.

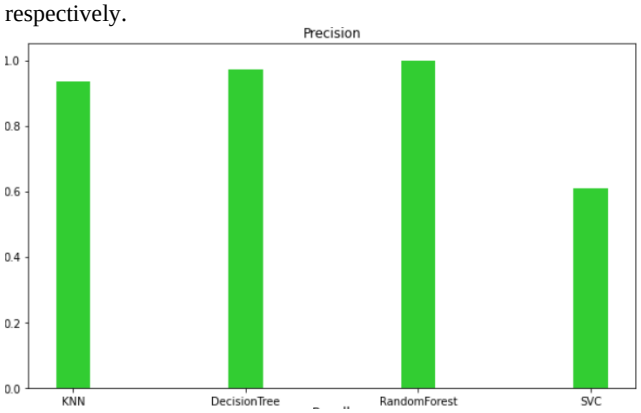

Fig 3: Bar chart of precision with respect to four classifiers

From the above figure 3, it can be inferred that, random forest classifier achieved better Precision value of 0.997 and Decision tree gave second highest of 0.983. While K-NN gave third highest of 0.890 and SVM gave least Precision value of 0.720 respectively.

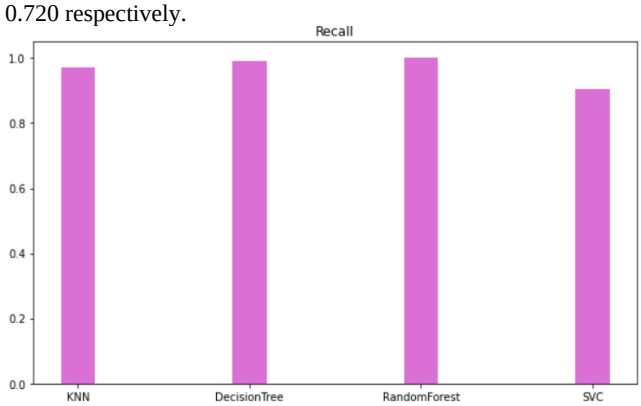

Fig 4: Bar chart of recall with respect to four classifiers

From the above figure 3, it can be inferred that, random forest classifier achieved better Recall value of 0.997 and Decision tree gave second highest of 0.985. While K-NN gave third highest of 0.930 and SVM gave least Recall value of 0.485 respectively.

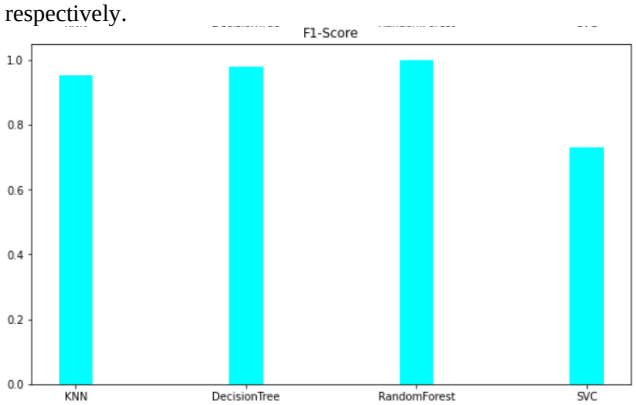

Fig5: Bar chart of F1 score with respect to four classifiers

From the above figure 4, it can be inferred that, random forest classifier achieved better F1score of 0.997 and Decision tree gave second highest of 0.987. While K-NN gave third highest of 0.909 and SVM gave least F1-Score value of 0.577 respectively.

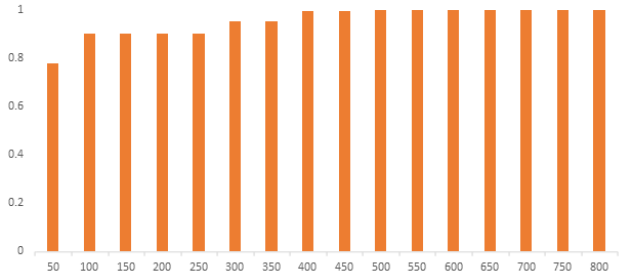

Fig6: Bar chart of Accuracy with respect to various iterations

From the above figure 11, it can be inferred that, based on various iterations at a range from 50 to 800 with an interval of 50, random forest is giving a continuous accuracy of 100% from interval 500 to 800 respectively. Hence, it can be taken that random forest gave best accuracy.

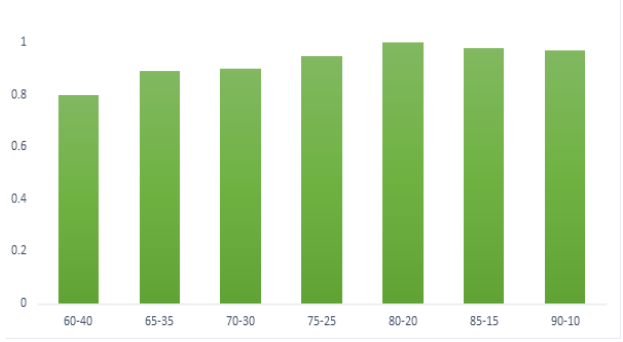

Fig 7: Bar chart of Accuracy with respect to various ratios

From the above figure 12, in can be inferred that, based on various comparisons of training and testing dataset ratios such 60-40, 65-35, 70-30, 80-20, 85-15 and 90-10 it can be seen that, at the ratio from 80-20 highest accuracy is attained.

5. CONCLUSION

As per the analysis of the overall result, the research was conducted on weather dataset by using Random Forest, Decision Tree, K-Nearest Neighbour and Support Vector Machine for detecting the dengue prevalence. Among all these algorithms, Random Forest gave better performance with respect to accuracy, precision, recall and F1-Score of 99.7% for detecting dengue prevalence. Till date, this research was conducted only on weather conditions and further it can be implemented by combining the blood profiles of the patients along with the weather conditions such as humidity, dewpoint, precipitation, maximum and minimum temperatures respectively.

6. REFERENCES

[1] C. P. G. Management, D. Infectionadults, and T. Edition, CPG Management of Dengue Infection In Adults (Third Edition) 2015 1. 2015.

[2] Wikipedia-http://www.bbc.com/news/world-asia-india-37415781.

[3] Rajathi, N., Brahanambika, R. and Manjubarkavi, K., 2018. Early detection of dengue using machine learning algorithms. International Journal of Pure and Applied Mathematics, [online] 118(18), pp.3881-3886. Available at: <: http://www.ijpam.eu> [Accessed 10 November 2021].

[4] Muzakki, M. and Nhita, F., 2018. The Spreading Prediction of Dengue Hemorrhagic Fever (DHF) in Bandung Regency Using K-Means Clustering and Support Vector Machine Algorithm. 2018 6th International Conference on Information and Communication Technology (ICoICT),.

[5] binti MohdZainee, N. and Chellappan, K., 2016. A preliminary dengue fever prediction model based on vital signs and blood profile. 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES),.

[6] Anitha, A. and Wise, D., 2018. Forecasting Dengue Fever using Classification Techniques in Data Mining. 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT),.

[7] Nakvisut, A. and Phienthrakul, T., 2018. Two-Step Prediction Technique for Dengue Outbreak in Thailand. 2018 International Electrical Engineering Congress (iEECON),.

[8] Fahmi, A., Purwitasari, D., Sumpeno, S. and Purnomo, M., 2020. Performance Evaluation of Classifiers for Predicting Infection Cases of Dengue Virus Based on Clinical Diagnosis Criteria. 2020 International Electronics Symposium (IES),.

[9] Husin N A, Salim N, Ahmad A R. Modeling of dengue outbreak prediction in Malaysia: A comparison of Neural Network and Nonlinear Regression Model[C]// International Symposium on Information Technology. IEEE, 2008, 1-4

[10] Mustaffa Z, Yusof Y. A Comparison of Normalization Techiques in Predicting Dengue Outbreak[J]. International Proceedings of Economics Development & Research, 2011.

[11] Quinlan, J. R. (1986). "Induction of decision trees" . Machine Learning. 1: 81–106. doi:10.1007/BF00116251. S2CID 189902138.

[12] Ho, T.K. (1995) Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 14-16 August 1995, 278-282.

[13] Bartlett P. and Shawe-Taylor J., "Generalization performance of support vector machine and other pattern classifiers", In C. ~Burges B. ~Scholkopf, editor, "Advances in Kernel Methods-Support Vector Learning", pp. 43–55, MIT press, 1998.

[14] Cover, Thomas M.; Hart, Peter E. (1967). "Nearest neighbor pattern classification" (PDF). IEEE Transactions on Information Theory. 13 (1): 21–27. CiteSeerX10.1.1.68.2616.doi:10.1109/TIT.1967.1053964