IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DENGUE DISEASE PREDICTION USING DATA MINING TECHNIQUES

[1]KHUSHBU ARVINDBHAI PATEL, [2]DR.MAHAMMADIDRISH I. SANDHI

[1]ASSISTANT PROFESSOR, [2]ASSOCIATE PROFESSOR& HEAD

[1]SHREE UTTAR GUJARAT BCA COLLEGE VEER NARMAD SOUTHGUJARAT UNIVERSITY, SURAT,

[2]SANKALCHAND PATEL COLLEGE OF ENGINEERING DEPARTMENT OF COMPUTER APPLICATION SANKALCHAND PATEL UNIVERSITY, VISNAGAR

Abstract: Data mining is used to extract useful information from large databases or data warehouses.[1] Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. For detecting a disease, the number of tests should be required from the patient. By using the data mining technique, the number of tests can be reduced.[2] Dengue is a threatening disease caused by Girl mosquitos.Dengue is caused by mosquito threatening the world now-a-days. If it is not curated on proper time, it will lead to death. World Health Organization (WHO) reported, dengue is prevalent in more than 81 countries all over the world.

Keyword: Dengue, Data mining, Prediction

1. INTRODUCTION

Data mining is a word of Computer Science which is sometimes it is also referred to as **knowledge discovery in databases** (KDD). Data mining means the process of selecting, discovering and modelling huge amounts of data. This process has become an increasingly subtle activity in medical science. [2]Dengue is the most viral disease caused by mosquito affecting humans. The virus is contracted from the bite of a striped Aedes aegypti mosquitoes that has previously bitten an infected person.Mosquito bite will intercommunicate the sickness. It will unfold from one person to a different person. The disease is caused by four serotypes of the infectious disease, a member of the genus RNA-flavivirus: DEN-1, DEN-2, DEN-3, and DEN-4.DEN-1 and DEN-2 serotypes most prevalent in India. Infection with the DEN virus many result in Dengue Fever (DF), Dengue Hemorrhagic fever (DHF) and Dengue Shock Syndrome (DSS).

[1]Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Volume 19, No 2

[2]Multi Disease Prediction using Data Mining Techniques K.Gomathi*, Dr. D. Shanmuga Priyaa**Article · December 2016

SYMPTOMS OF DENGUE:

- The sick individual might be influenced extreme migraine and high fever
- While moving the eyes, an intense pain is suffered behind the eyes by the patient.
- There is an agony in joints in an influenced person.
- Bone and muscle torments square measure another normal side effect of dengue.
- The rashes might have appeared in a diseased person.
- At times gentle draining is likewise recognizable by a person.

The dengue fever is classified based on the various symptoms. In the initial stages, the dengue fever is difficult to be identified. Various data mining techniques have been given by various researchers to assess the intensity of dengue fever.

2. REVIEW OF LITERATURE

a. K means clustering algorithm for dengue fever assessment was provided by P Manivannam and Dr. P Isakki. It is a technique to classify or bifurcate various attributes of the objects into K number of groups. Through, K means clustering, work was done for predicting dengue fever based on the age group categorised data. This technique is suitable for predicting dengue fever patients with serotypes.[3]

b. M Mufli Muzakki et al proposed prediction of dengue DHF in Bandung Regency through k-Means clustering and Support Vector Machine (SVM) algorithm. These data are used from Climatological, Geophysical and Meteorological agency in Bandung Regency. It was concluded that weather data could be used to predict DHF disease because of the direct relationship between weather and DHF. The recommendations of this research work could be useful for Health Department of Bandung Regency and to increase the awareness of people about DHF disease.[4]

c. Abdul Mahatir Najar and Mohammad Isa Irawan et al conducted research on dengue hemorrhagic fever. DHF is transmitted through Aedes Aegypti and Aedes Albopictus mosquitos. DHF was commonly found in tropical and sub-tropical areas. It was found that machine learning can predict the risk level of DHF through 50 neurons.[5]

d. Iwan Inrawan Wiratmadja, Siti Yaumi Salamah et al worked on predicting length of stay in hospital. Accuracy of 71.57 % was achieved through decision tree. Through decision tree technique, prototype PF dengue patient's length of stay prediction system was developed.[6]

[3]P. Manivannan, Dr.P.Isakki, Dengue Fever Prediction Using K-Means Clustering Algorithm, IEEE International Conference On Intelligent Techniques Incontrol,Optimization And Signal Processing, 978-1-5090-4778-9/17/©2017

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

[5]Abdul mahatir najar,Mohammad isa irawan et al, Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction,IEEE International Conference On Smart Computing And Electronic Enterprise (ICSCEE2018) ©2018

[6]Iwan Inrawan Wiratmadja, Siti Yaumi Salamah & Rajesri Govindaraju, Journal of engineering and Technology Science., Vol. 50, No.1, 2018, 110-126

3. OBJECTIVES

The research is aimed at using a few classifying techniques to predict the dengue affected cases in Surat district and surrounding areas geographically. An attempt is also made to compare different classification algorithms by using graphs, and dataset. I have implemented some techniques for data analysis by using SPSS Modeler tool.

4. CLASSIFICATION

In data mining, the classification refers to recognising and detectingfeatures of infection among patients and forecast the techniques which is the best on the basis of WEKA analysis.

In this paper, five techniques of data mining have been used for the purpose to be achieved. These techniques uses Explorer interface and it depends on dissimilar techniques NB, J48, RT, LMT and SMO.

All these techniques which are used have been applied on Dengue data set. In this analysis, classification and accuracy have been observed. (Table 1).

Attribute Name	**Definition**
Correctly Classified	Shows **the** percentage of correctness of **how many instan**ces are categorized accurately in the test.
Incorrectly classified	It indicates **the percen**tage of incorrectness of **how many instances are categorized** inaccurately in the test.
TP Rate	It shows the rate of true positives i.e. attributes which are correctly classified.
FP Rate	It indicates the rate of false negatives. i.e. instances which are negative but are classified as true.
ROC Rate	It is the method of **visualizing, organizing and selecting classifiers based on their performance.** It is helpful in the signal detection.
Precision	It is easy to calculate and made precision based on that.It indicates the percentage of relevant results in your dataset.
Types **of Precision** **TN** **TP** **FN** **FP**	Four types of precision rate are as under: Predicted was negative when the variables are negative Prediction was positive when the variables were positive Prediction was negative when the variables were positive. Prediction was positive when the variables were negative.
Accuracy	It is the propensity of how well a given predictor can guess the value of predicted attribute for a new **data**.
Error Rate	Error rate is due to selection of property which is not suitable for classification. Error Rate=1 - Accuracy

[Table: 1 Attributes definition]

5. DATASET:

Collection of data refers to the data set. The data set deals with the contents of a single database table or data matrix, in which all the variables in each column corresponds to all the variables in each row of the same table. In this paper, dataset of 115 has been taken for the research. This dataset was taken from Shree General Hospital Surat.

For the testing and analysing data set accepted, WEKA tool has been used in this research. To measure the accuracy, some data were classified and the rest were tested.

Age	Gender	Date	Headache	High fever	Joints pain	Muscle pains	Rashes	Vomiting	Bleeding	Dengue
42	Female	19-01-2019	No	Yes	No	Yes	No	No	No	Negative
18	Male	12-04-2019	Yes	Yes	Yes	No	No	No	No	Positive
26	Female	14-06-2019	No	Yes	Yes	Yes	No	No	Yes	Positive
32	Male	17-06-2019	Yes	Yes	No	No	No	No	No	Negative
40	Female	21-06-2019	Yes	Yes	Yes	Yes	Yes	No	No	Positive
39	Female	03-07-2019	No	Yes	No	Yes	Yes	Yes	No	Positive
36	Female	03-07-2019	Yes	Yes	Yes	No	No	No	No	Positive
35	Male	23-07-2019	No	Yes	No	Yes	No	No	No	Negative
30	Male	25-07-2019	No	Yes	No	Yes	No	Yes	No	Negative
15	Male	29-07-2019	No	Yes	No	Yes	Yes	Yes	No	Positive
36	Male	02-08-2019	Yes	Yes	No	No	Yes	No	Yes	Positive
20	Female	05-08-2019	Yes	Yes	No	Yes	No	No	Yes	Positive
25	Female	07-08-2019	No	Yes	Yes	Yes	Yes	No	No	Positive
20	Male	09-08-2019	Yes	Yes	Yes	No	No	No	No	Positive
26	Female	09-08-2019	No	Yes	Yes	Yes	No	No	Yes	Positive
35	Male	10-08-2019	Yes	Yes	Yes	Yes	Yes	No	No	Positive
21	Female	17-08-2019	No	Yes	No	Yes	Yes	Yes	No	Positive
25	Male	18-08-2019	Yes	Yes	Yes	No	No	No	No	Positive
35	Female	20-08-2019	No	Yes	No	Yes	No	No	No	Negative
36	Male	26-08-2019	No	Yes	No	Yes	Yes	Yes	No	Positive

[Figure. 1: Dataset]

6. ATTRIBUTES:

In WEKA, CSV file has been used. The Attributes that we have chosen for the testing of dengue are Headache, High Fever, Joints Pain, Muscle Pains, Rashes, Vomiting, Bleeding and other indications with class label of Dengue with Positive and Negative Consequences (Fig.2) The attributes description is given in (Table 2).

Attributes	Description
PID	Id of Patient
Age	Age of Patient
Gender	Gender of Patient
Date	Fever Date
Headache	Yes or No
High fever	Yes or No
Joints pain	Yes or No
Muscle pains	Yes or No
Rashes	Yes or No
Vomiting	Yes or No
Bleeding	Yes or No
Dengue	Positive or Negative

[Table: 2 Attributes Description]

7. DATA MINING TECHNIQUES:

For predicting Dengue virus, different data mining techniques have been used. By using different DM techniques, predictions have been done for the classification and accuracy of Dengue fever. . Accuracy can be observed by selecting the following procedures: NB, J48, RT, LMT and SMO.

The techniques we are using are following:

- NB
- J48
- RT
- LMT
- SMO

1. NAIVE BAYES RULE:

Naïve Bayes is used for arithmetical prediction, for forecasting class attributes possibility. This prediction is based on the simple Bayes formula. With NB classifier, we can compare the performance of ID3 and selected neural classifiers. We have verified the data set on WEKA tool, and got the results which is mentioned in the (Table 3)

2. J48:

Ross Quinlan has developed the technique of C4.5 to generate a decision tree. It is enlargement of Quinlan's earlier ID3 technique. The decision tree created by C4.5 techniques can also be used for the classification of dataset. This technique construct the decision tree same as ID3, from the set of training data set. We tested out the training data on WEKA tool with J48 technique and conclude the outcomes in the table 6.

3. RT:

Random Tree creates a decision tree from randomly selected subset of training set. It uses some ideas to create random data set to build an ID3 (Figure 4). In standard tree, using the best split among all variables, each node is divided. In the Random Forest, every node is split through the best among the subset of predictors chosen at that particular node.[7]We tested Dengue prediction through WEKA tool and data presented in the table 5.

4. LMT:

Logistic Model Tree (LMT), which combines Standard Decision Tree (DT) and Linear Logistic Regression algorithm in a single tree. LMT was applied in this research paper and the concluding results were bifurcated in the table. Considering the allocation of data set, it was also estimated that the LMT algorithm produced the most accurate results.[8]

5. SMO:

SMO (Sequential Minimal Optimisation) is used for training support vector machines which was invented by John Platt in 1988. SVM requires a very large number of Quadratic Programming problems which are split into a series of smallest QP problems by SMO.[9]

To predict the occurrence of each Dengue data set, we assessed the output of classifier by altered measurements. Through SMO in WEKA, we got the result which was depicted in the table 7.

[7]R.Sanjudevi1, D. Savitha2, DENGUE FEVER PREDICTION USING CLASSIFICATION TECHNIQUES, Journal of International Research Journal of Engineering and Technology, Volume: 06 Issue: 02 | Feb 2019

[8]Niels Landweh, Mark Hall and Eibe Frank,"Logistic Model Trees",extended version of a paper that appeared in the Proceedings of the 14th European Conference on Machine Learning (Landwehr et al., 2003)

[9]Keerthi SS, et.al. (2001) Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation 13: 637-649.

8. METHODOLOGY

Below are the process steps:

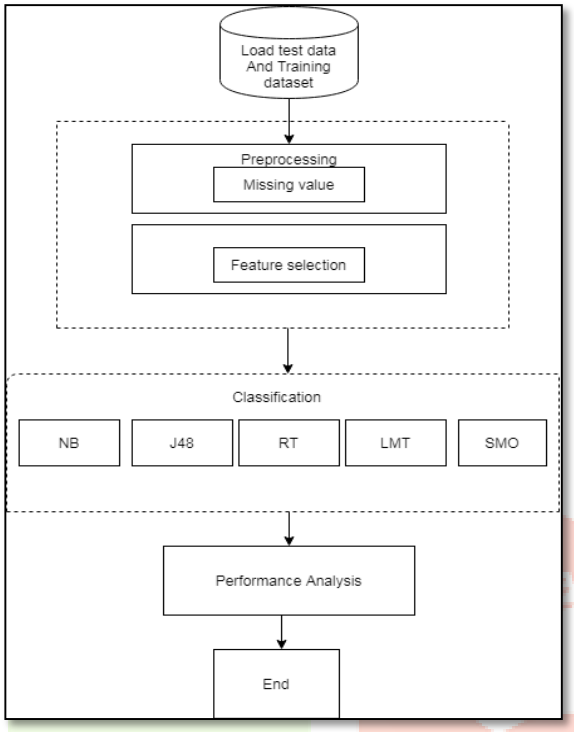
Step 1: Loading of Test data and dataset.

Step 2: Pre-Processing of Dataset through missing value imputation technique.

Step 3: Through forward and backward selection method, feature selection was undertaken.

Step 4: In order to predict Dengue, the classification algorithm was used.

Step 5: accuracy was measured through the classification algorithm based on the results achieved.[10]



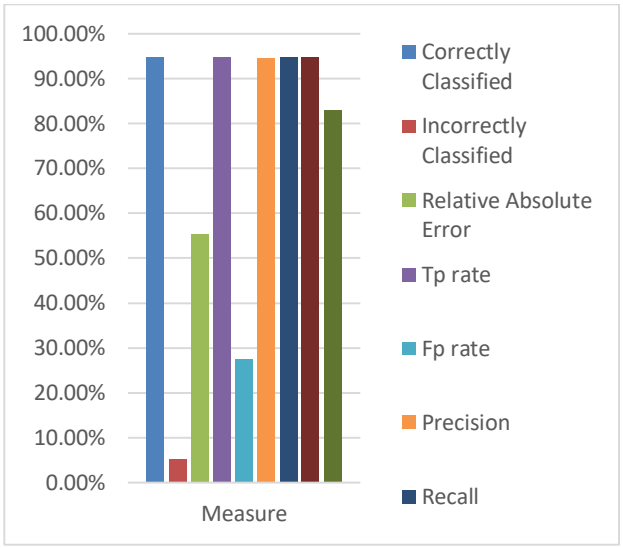
8.1.IMPLEMENTATION TOOL WEKA

University of Waikato invented WEKA tool for data mining which is mostly used tool. WEKA is widely being used for developing Machine Learning and applications for data mining problems. It is a combination of machine learning algorithms which is used for data mining tasks. Machine learning process can be directly applied from the explorer menu on the dataset. It includes various processes such as clustering, classification, regression, Decision making tree and so on. It is widely being used tool as it is open source and free of cost.

[10]Kamran Shaukat, Nayyer Masood, Sundas Mehreen, Ulya Azmeen. (2015).
Dengue Fever Prediction: A Data Mining Problem, Journal of Data Mining in Genomics & Proteomics, Volume 6, Issue 3.

Attributes	Measure
Correctly Classified	94.78%
Incorrectly Classified	5.22%
Relative Absolute Error	55.26%
TP rate	0.948
FP rate	0.275
Precision	0.945
Recall	0.948
F-measure	0.946
Roc Area	0.83

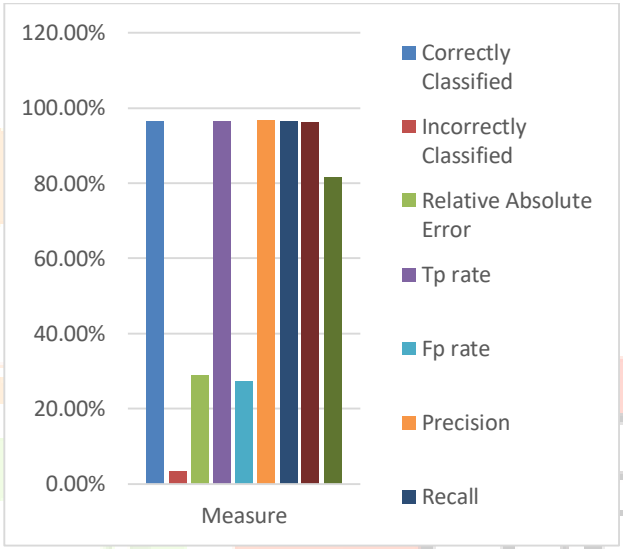
[Table: 3 Naive Bayes Technique]



[Figure: 2 Naive Bayes Chart]

Attributes	Measure
Correctly Classified	96.52%
Incorrectly Classified	3.48%
Relative Absolute Error	28.85%
TP rate	0.965
FP rate	0.273
Precision	0.967
Recall	0.965
F-measure	0.962
Roc Area	0.817

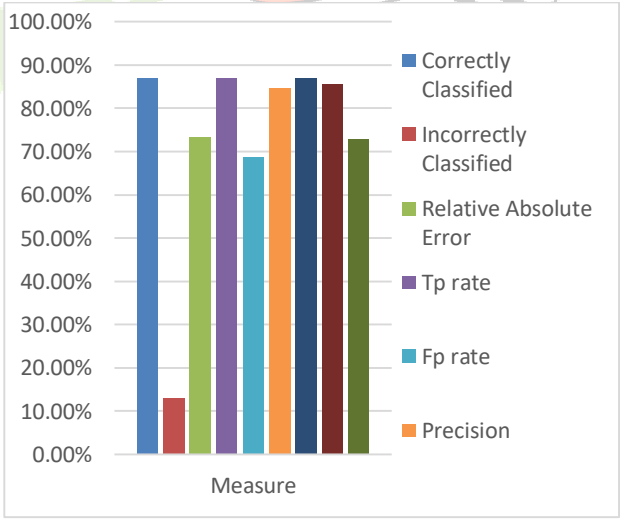
[Table: 4 J48 Technique]



[Figure: 3 J48 Chart]

Attributes	Measure
Correctly Classified	86.96%
Incorrectly Classified	13.04%
Relative Absolute Error	73.39%
TP rate	0.87
FP rate	0.688
Precision	0.846
Recall	0.87
F-measure	0.856
Roc Area	0.729

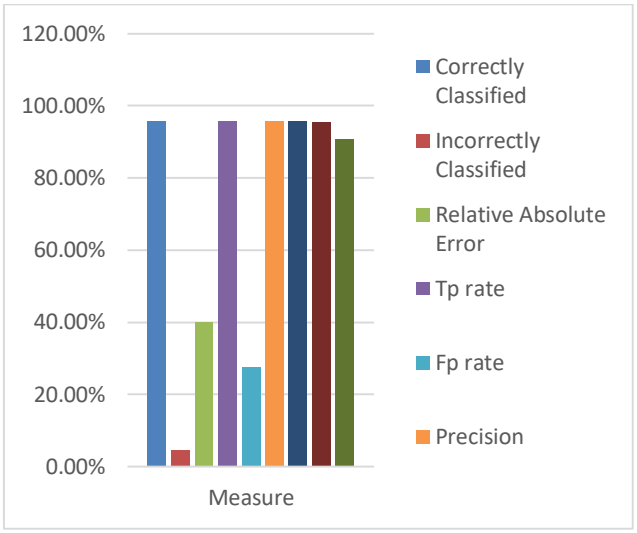
[Table: 5 Random Tree Technique]



[Figure :4 Random Tree Chart]

Attributes	Measure
Correctly Classified	95.65%
Incorrectly Classified	4.35%
Relative Absolute Error	39.93%
TP rate	0.957
FP rate	0.274
Precision	0.955
Recall	0.957
F-measure	0.954
Roc Area	0.908

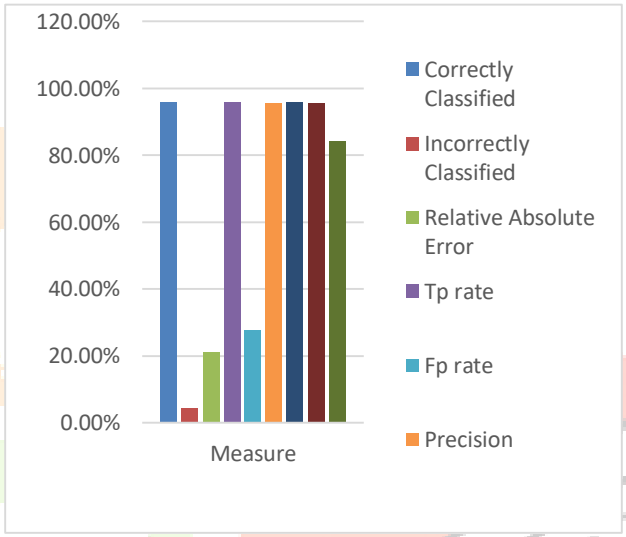
[Table :6 LMT Technique]



[Figure :5 LMT Chart]

Attributes	Measure
Correctly Classified	95.65%
Incorrectly Classified	4.35%
Relative Absolute Error	21.06%
TP rate	0.957
FP rate	0.274
Precision	0.955
Recall	0.957
F-measure	0.954
Roc Area	0.841

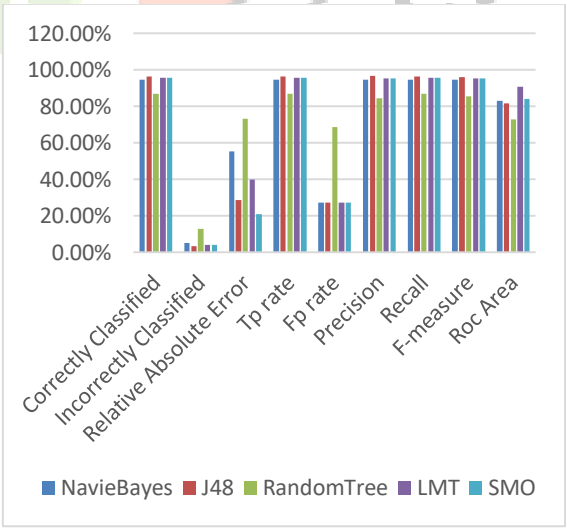
[Table : 7 SMO Technique]



[Figure : 6 SMO Chart]

Techniques	TP Rate	ROC Rate	Error Rate	Accuracy
NavieBayes	0.948	0.83	0.052	0.948
J48	0.965	0.817	0.035	0.965
Random Tree	0.87	0.729	0.13	0.87
LMT	0.957	0.908	0.043	0.957
SMO	0.957	0.841	0.043	0.957

[Table : 8 Comparison Table]



[Figure : 7 Comparison Chart]

COMPARISON

Results have been analysed by us through the usage of five aforesaid techniques of classification. When we have done the comparison among all of them we concluded that J48 Technique is greatest among all others. As the accuracy of J48 is 96.58% which was biggest of all. J48 is the best also for the aim that it gives the probability and efficiency. Given below is the comparison of all the techniques (Table 8). The graph comparison is given in (Figure 7).

CONCLUSION

Prediction of dengue disease using WEKA Data Mining tool is the sole aim of this research. Actually it has four edges out of which we have used only one edge which is Explorer. In these five techniques of classification, i.e., NB, J48, RT, LMT and SMO. Using Weka Data Mining tool to evaluate the accuracy, these techniques were applied. Accuracy was compared after testing the results through aforesaid techniques. Classifier accuracy was compared with each other on based on correctly classified instances, a precision, error rate, TP rate, FP rate and ROC Area. Over Explorer technique it has concluded that J48 is the top performance classifier techniques by way that, they has achieved an accuracy of 96.58%, takes fewer time to run and shows ROC area=0.035, and had smallest error rate.

REFERENCES:

[1] HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Volume 19, No 2.

[2]Multi Disease Prediction using Data Mining Techniques K.Gomathi*, Dr. D. Shanmuga Priyaa**Article · December 2016

[3] P. Manivannan, Dr.P.Isakki, Dengue Fever Prediction Using K-Means Clustering Algorithm, IEEE INTERNATIONAL CONFERENCE ON INTELLIGENT TECHNIQUES INCONTROL,OPTIMIZATION AND SIGNAL PROCESSING, 978-1-5090-4778-9/17/©2017

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

[5] Abdul mahatir najar,Mohammad isa irawan et al, Extreme Learning Machine Method for Dengue Hemorrhagic Fever Outbreak Risk Level Prediction,IEEE International Conference On Smart Computing And Electronic Enterprise (ICSCEE2018) ©2018

[6] Iwan Inrawan Wiratmadja, Siti Yaumi Salamah & Rajesri Govindaraju, Journal of engineering and Technology Science., Vol. 50, No.1, 2018, 110-126

[7]R.Sanjudevi1, D. Savitha2, DENGUE FEVER PREDICTION USING CLASSIFICATION TECHNIQUES, Journal of International Research Journal of Engineering and Technology, Volume: 06 Issue: 02 | Feb 2019

[8] Niels Landweh, Mark Hall and Eibe Frank,"Logistic Model Trees",extended version of a paper that appeared in the Proceedings of the 14th European Conference on Machine Learning (Landwehr et al., 2003)

[9] Keerthi SS, et.al. (2001) Improvements to Platt‘s SMO algorithm for SVM classifier design. Neural Computation 13: 637-649.

[10]Kamran Shaukat, Nayyer Masood, Sundas Mehreen, Ulya Azmeen. (2015).
Dengue Fever Prediction: A Data Mining Problem,Journal of Data Mining in Genomics & Proteomics, Volume 6, Issue 3.