Evaluating The Performance of Machine Learning using Feature Selection Methods on Dengue Dataset

Subhram Dasgupta, Naman Sharma, Sweta Sinha, Raghavendra S

Abstract: ***Dengue fever is a mosquito-borne disease transmitted by the bite of an Aedes mosquito infected with a dengue virus. The bites of an infected female Aedes mosquito which gets the virus while feeding on the infected persons blood, transmits the virus to others. Dengue transmission is climate sensitive for several reasons such as temperature, humidity, rainfall, etc. Areas having higher vapor pressure and rainfall rate are most vulnerable to the spreading of the dengue disease. So to find the important features responsible for spreading the dengue we have used the classification algorithms. Machine learning is one of the key methods used in modern day analysis. Many algorithms have been used for medical purposes. Dengue disease is one of the serious contagious diseases. To find the features related to spreading of dengue disease, we have used popular machine learning algorithms. This proposed work focuses on evaluating the performances of the various machine learning techniques like- Random Forest Classifier (RFC), Decision Tree Classifier (DTC) and Linear Support Vector Machine (LSVM). Predictive Mean Matching is applied for preprocessing of the data and percentage split is applied for resampling of the data. Information gain values for each of the attributes are calculated. The attributes are sorted on the basis of information gain values. Feature selection methods (FSMs) such as Forward Selection (FS) and Backward Elimination (BE) are applied to choose the finest subset of the attributes, so that the algorithm runs more efficiently with a lower run time. It also results in the improvement of the accuracy. The attributes selected by the Feature Selection Methods are the main attributes which results in the probable effects of global weather change on human healthiness.***

Keywords: Logistic regression; Artificial neural network; Random forest; Support vector machine; Neural network with 10-fold, dengue disease, Neural network, R.

I. INTRODUCTION

Machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the dengue dataset as input given to it and then uses this learning to classify new observation. It has categories of algorithms that allow software application to handle new situation via analysis, self-training, observation and experience. A “training data” which is a mathematical design of sample data, is built to make prediction without programming to accomplish the required task.

There are many algorithms in machine learning like DTC, linear regression, RFC and LSVM. The algorithms used in this proposed work are RFC, LSVM and DTC. To determine the best model, we have performed different conditions for predicting the accuracy, which includes elimination method with percentage split as test option. The process of data mining is performed using RStudio. RStudio is open source software that provides a collection of machine learning and data mining algorithm for data preprocessing, classification, regression, clustering, association rules and visualization. In this process, dengue dataset is embedded into RStudio tool and results are predicted using machine learning algorithms. Dengue dataset consists of various attributes such as humidity, trees, temperature, etc. After collecting dataset of dengue, preprocessing of data is performed to handle missing data. This data obtained after preprocessing is used to calculate Information gain (Ig) value of each attributes. This process is known as Entropy evaluation method. To improvise selection of appropriate and important attributes we use FSMs methods such as FS and BE which also help in reducing complexity and improve the accuracy of model. In this proposed work we try to evaluate the performance of classification algorithms like LSVM, DTC and RFC on dengue dataset using FSM’s like FS and BE with percentage split as test option. Several subsets are constructed based on the Ig values of each attributes. Each of the subsets are trained to perform classification and the accuracy of the model is noted. The main objective of the proposed work is to improvise the efficiency of the machine learning processes which are used to forecast the accuracy of the classification models. To improvise the efficiency of algorithms, FSM’s are used. This permits the machine learning process to train faster, it also reduces the amount of attributes to improve the accuracy of a model if the right subset is chosen. So, we used FSM’s to develop an accurate predictive model for machine learning algorithm.

The main limitations of this work is in:

- Finding an appropriate dataset related to dengue disease for the implementation of the algorithms.
- The problem of handling the missing values is one of the important challenge in finding the prediction accuracy of Machine learning techniques.
- Time constraint is another important challenge.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)

Dr. Raghavendra S*, Associate Professor, Department of Computer Science and Engineering at CHRIST DEEMED TO BE UNIVERSITY, Bangalore. (Karnataka), India.

Subhram Dasgupta, Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore (Karnataka), India.

Naman Sharma, Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore (Karnataka), India.

Sweta Sinha, Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore (Karnataka), India.

II. LITERATURE SURVEY

This proposed work is based on classification algorithm which is used to evaluate dengue dataset.The author uses spatial auto correlation to predicting the accuracy of transmitting the dengue fever. Based on long term average vapor pressure an accuracy of dengue fever transmission accuracy was 89%, which means that climate change is the major factor for spreading of vector-borne disease [1]. The author notes the transmission of disease in different area. A team was built which performed survey in selected areas of Australia. The value of adaptation measures to diminish the hazard of damage from upcoming climate modification, and from present-day climate variability, was recognized in this. The conclusion obtained was climate change such as: humidity, temperature, rate of rainfall, etc. are major reason of transmission of diseases [2]. A global model of malaria transmission was developed to estimate the potential impact of climate change on seasonal transmission and population at risk of the disease. Risk indicates about the weather conditions of that area. Climate change in East Africa, Central Asia, China and areas of South America where monitored. The author has described a new method for describing vulnerability to the potential impacts of climate change [3].

Machine learning classification algorithms are used to detect the presence of dengue in a particular patient. It uses linear regression, decision trees and SVM to find out the accuracy, so that it can compare the accuracy ratio and find out the best model for dengue detection [4]. The author uses various machine learning algorithms for predicting dengue disease. First identify the symptoms of dengue in patients and prediction begins from this identification. The data sets are used for classification and to predict the accuracy. Decision tree was used to predict the chances of occurrences of dengue diseases in a tribal area. The database is analyzed for the creation of an unsupervised model to identify the most significant parameters of affected area and to predict the chances of hitting the disease using the supervised classifier model [5]. The authors proposed a disease prediction system using Random Forest Algorithm (RFA). The key intention is to predict the disease which input symptoms is taken from patient or user. Random forest algorithm maintains best accuracy compared with others. After outcome forecast the disease, reference system will work on their predicted disease [6]. A prediction that integrates Least Squares Support Vector Machines (LS-SVM) in forecasting forthcoming dengue outbreak. The work presented in this incorporates the Least Squares Support Vector Machines in predicting dengue outbreak for five districts in Selangor, Malaysia. Prediction accuracy obtained in the undertaken experiment is compared against the one implemented using Artificial Neural Network model [7]. Feature selection and variable are the main focus of many research in domains of application where datasets comprising of tens or hundreds of thousands of variables are available. The main task of variable choice is 3-fold: refining the predictors achievements, furnishing quicker and cost-effective predictors, and furnishing a improved understanding of the fundamental process that produced the data. The contributions of this special issue cover a wide variety of features of similar problems: furnishing an improved meaning of the objective purpose, feature building, feature status, multivariate feature range, effective search approaches, and feature validity assessment procedures [8]. The dengue viruses occur in 4 serotypes (DENV-1 to DENV-4). Dengue disease varies from minor feverish disease to serious hemorrhagic illness. Predicting the relationship between the dengue serotypes will confidently support the bioinformaticians and biotechnologists to change one stage onward to notice antibiotic for dengue. The author focuses 4 stages namely attribute selection, preprocessing, grouping and forecasting the dengue illness. For pre-processing R 3.3.2 tool is employed for the household of dengue dataset. D win’s technique was used to produce full dataset by replacing all lost values for insignificant and numerical qualities with style and mean value. Dengue virus can be foreseen by using dissimilar data mining methods. The main aim of research work is to forecast the people affected by dengue liable upon classification of age group using K-means clustering algorithm has been implemented [9]. A creative data mining procedure for foreseeing the dengue through medicinal histories of patients is anticipated. Dengue is an very regular disease these days in all peoples and in all age crowds. So pulling out the dengue information in a creative means is a elementary issue. The altered J48 classifier is exploited to figure the accuracy degree of the data mining system. The WEKA was exploited for generating the altered J48 classifiers. Exploratory consequences confirmed a notable modification over the existing J48 algorithm [10]. Medical data contain very valuable information which can save many lives if it is analyzed and utilized efficiently. Efficient analysis of this large volume of data demands the right choice of predictors and this in turn can impact the accuracy of the decision support system. Dimensionality reduction and feature subset selection are two techniques to reduce the number of features used in classification. In this an empirical evaluation of four feature selection methods when applied in conjunction with RFC. The feature selection techniques applied are Relief feature selection algorithm, Random forest selector, Recursive feature elimination and Boruta Feature selection algorithm. Results show that feature selection methods boosts the performance of the classifiers and in this case the features selected by the Boruta feature selection algorithm gives the best results [11]. From the literature survey we can see that the existing model on potential effects of climate change on transmission of vector borne diseases is with the accuracy of 89% which is based on long-term average vapor pressure. The accuracy can be improved further and more important attributes is to be identified. The proposed work uses RFC, DTC and LSVM model for prediction. And to improve the process of selecting attributes from dataset, FSM’s are used. FSM’s includes FS and BE. As test options we use percentage split and the performance of the design is evaluated using classification accuracy.

III. PROPOSED FRAMEWORK

The dataset was taken based upon the information whether dengue disease was recorded between 1961 and 1990. The dataset contains 2000 observations with 13 attributes. The data is collected from the administrative regions spread across the world. The information provided on climate and tree covers were given for each half degree of latitude by longitude.

The decision variable (No / Yes) was given by the administrative region. The attributes such as humidity, temperature, tree cover data are given in 50/90 percentiles, where the percentiles were calculated across the pixels over the mentioned administrative region. The dataset which is loaded in the RStudio software is in the (CSV) format. The data is distributed into (2000 Rows * 13 Columns) a total of 26,000 cells. The data is preprocessed to remove the unwanted data (NA's). The data is then split using the percentage split method to form training sets to train the data and a test set to evaluate it. In the percentage split method the whole data set after preprocessing is split into training set and testing set on the basis of the values entered by the programmer, We chose to take 4 different percentage splits (50 / 50, 60 / 40, 75 / 25, 80 / 20). The algorithms implemented to predict the accuracy are LSVM, RFC and DTC. The selection of the algorithms is based upon the study and the best fitted algorithms are taken to find the maximum possible accuracy for the dataset. LSVM is selected because it can be used for both classification and regression and can be applied to both linear and nonlinear data. It is a supervised machine learning algorithm. Unlike LSVM, DTC and RFC also serve the same purpose. These algorithms are used to calculate the accuracy. The Ig values of the attributes were calculated and the data attributes 12 were accordingly sorted for applying the FSM. The FS and BE methods are applied. The prime objective was to reduce the training time and also to improve the accuracy predicted from the earlier models. The results from the model was applied to make the projections for the possibility of the dengue fever in future. It is also done to predict the deciding factors for the spreading of the vector borne disease.

The dataset used for this proposed work contained incomplete data which were removed so that the classification algorithm runs with proper data input. 'pmm' is used to replace the incomplete data with the estimated values. The following are the different methods implemented in the proposed work :

1. LSVM: A Support Vector Machine (SVM) is a classifier defined by a separating hyperplane which can make distinct observations within a dataset. Given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes the data into new sets with similar attributes. In two dimensional space this hyperplane is a line dividing a plane in two parts where the class attribute responses lay on the either sides of the line.
2. DTC: DTC is a systematic approach which repetitively divides the dataset into sub parts by identifying lines. It organizes a series of test questions and conditions in a tree structure. In decision tree, the root and the internal nodes contains the test conditions to separate the dataset having different attributes. The terminal nodes are assigned the values (YES/NO). Once the tree is constructed the test conditions are applied, based on that the appropriate branch is selected. The iterative process terminates when the criteria's of the classifier attributes are met or the classes which are subdivided are pure(i.e. the condition when the divided classes are having absolute distinct values).
3. RFC: RFC is an ensemble algorithm. Ensembled algorithms combines more than one algorithms of same or different kind for classifying objects. It creates a set of decision trees from a randomly selected subsets of the training sets. Due to the combinations it formulates, the accuracy RFC produces is much better than the other classification methods.
4. FSMs: FSMs are basically used for the selection of variables which will reduce the run time of the algorithms and also increase the accuracy. It is performed to increase the efficiency of the algorithm. The main advantages are:
   - The machine learning algorithms are able to train faster.
   - Reduces the complexity of the model making it easier to understand.
   - Accuracy is improved on applying the right subset and also helps to reduce overfitting.
5. Wrapper Methods: In this method, a subset of the features are used to train the model. Based on the results of the previous model, the addition or removal of the features from the dataset is decided. The two methodologies implemented are:
   A. FS: It is an iterative model. In FS method we start with having no feature. The attributes are then consecutively added till the best result for a combination of attributes is achieved. Its purpose is to improve the performance of the model.
   B. BE: In this method, we start with having all the features in the model and then subsequently deleting the attributes on the basis of the information gain values until the best combination for the model is achieved. This process is terminated when there is no improvement on the removal of the features.

The architecture followed in the proposed work is shown in Figure 1.

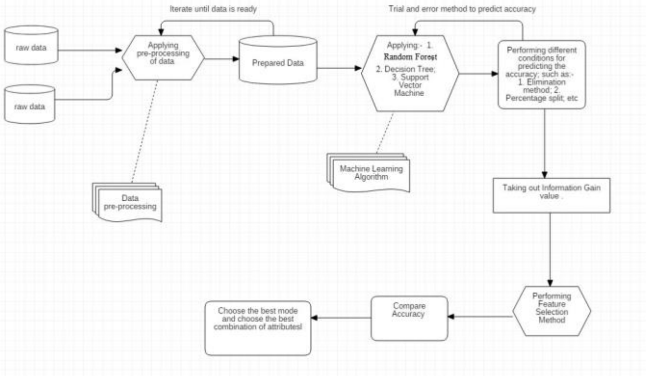

Iterate until data is ready
Trial and error method to predict accuracy
raw data
raw data
Applying pre-processing of data
Prepared Data
Applying:- 1. Random Forest 2. Decision Tree; 3. Support Vector Machine
Performing different conditions for predicting the accuracy; such as:- 1. Elimination method; 2. Percentage split; etc
Data pre-processing
Machine Learning Algorithm
Taking out Information Gain value .
Performing Feature Selection Method
Compare Accuracy
Choose the best mode and choose the best combination of attributes!

Fig. 1. Proposed Model for Prediction

The flow of work in the proposed work is listed below:

Step 1: The raw dataset is gathered.

Step 2: The raw dataset is pre-processed to eliminate all the unwanted data so that the result of the algorithms are not affected.

Step 3: The RFC, DTC and Linear SVM algorithms are implemented on the prepared data to calculate the accuracy of CA's.

Step 4: The prepared data is labelled into training and testing data using the percentage split method or the cross validation methods.

Step 5: The Ig values of the attributes are calculated to rearrange the data so that the wrapper methods of the feature selection can be applied.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Step 6: After performing the FSM's the accuracy is compared and the best subset is chosen so that the best combination of the attributes can be chosen which gives more better accuracy than the previous observation.

IV. RESULTS AND DISCUSSION

The Ig value of each attribute of the dengue dataset is shown in Table 1.

TABLE 1: Information Gain values of each attributes of dengue dataset

Attribute	Information Gain Value
h10pix90	0.4283218
h10pix	0.4153768
humid	0.3537384
humid90	0.3535968
Ymax	0.3490389
Ymin	0.3433688
temp90	0.2917194
temp	0.2763417
Xmax	0.202118
Xmin	0.1825686
trees	0.0408454
trees90	0.0371557

Table 2 and Table 3 shows the attributes arranged in ascending order and descending order of their Ig values respectively.

Table 2: Attributes in ascending order of their Ig value

Attribute	Information Gain Value
trees90	0.0371557
trees	0.0408454
Xmin	0.1825686
Xmax	0.202118
temp	0.2763417
temp90	0.2917194
Ymin	0.3436879
Ymax	0.349039
humid90	0.3535968
humid	0.3537836
h10pix	0.4153768
h10pix90	0.4283218

Table 3: Attributes in descending order of their Ig value

Attribute	Information Gain Value
h10pix90	0.428322
h10pix	0.415377
humid	0.353738
humid90	0.353597
Ymax	0.349039
Ymin	0.343369
temp90	0.291719
temp	0.276342
Xmax	0.202118
Xmin	0.182569
Trees	0.040845
trees90	0.037156

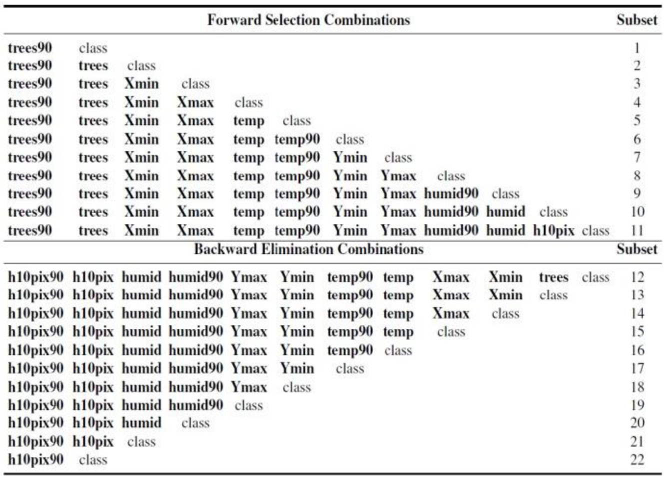
Table 4 shows the accuracy achieved by all the three methods without FSMs for different split ratio.

Table 4: Accuracy of all the three methods without using FSMs

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9367521	0.9457143	0.9524362	0.9545455
DTC	0.9096757	0.9152778	0.9186857	0.9229112
LSVM	0.914	0.915	0.926	0.9156

Table 5 shows the different combinations of attributes obtained after using FS and BE based on the Ig values.

Table 5: Different combination of attributes



Forward Selection Combinations												Subset
trees90	class											1
trees90	trees	class										2
trees90	trees	Xmin	class									3
trees90	trees	Xmin	Xmax	class								4
trees90	trees	Xmin	Xmax	temp	class							5
trees90	trees	Xmin	Xmax	temp	temp90	class						6
trees90	trees	Xmin	Xmax	temp	temp90	Ymin	class					7
trees90	trees	Xmin	Xmax	temp	temp90	Ymin	Ymax	class				8
trees90	trees	Xmin	Xmax	temp	temp90	Ymin	Ymax	humid90	class			9
trees90	trees	Xmin	Xmax	temp	temp90	Ymin	Ymax	humid90	humid	class		10
trees90	trees	Xmin	Xmax	temp	temp90	Ymin	Ymax	humid90	humid	h10pix	class	11
Backward Elimination Combinations												**Subset**
h10pix90	h10pix	humid	humid90	Ymax	Ymin	temp90	temp	Xmax	Xmin	trees	class	12
h10pix90	h10pix	humid	humid90	Ymax	Ymin	temp90	temp	Xmax	Xmin	class		13
h10pix90	h10pix	humid	humid90	Ymax	Ymin	temp90	temp	Xmax	class			14
h10pix90	h10pix	humid	humid90	Ymax	Ymin	temp90	temp	class				15
h10pix90	h10pix	humid	humid90	Ymax	Ymin	temp90	class					16
h10pix90	h10pix	humid	humid90	Ymax	Ymin	class						17
h10pix90	h10pix	humid	humid90	Ymax	class							18
h10pix90	h10pix	humid	humid90	class								19
h10pix90	h10pix	humid	class									20
h10pix90	h10pix	class										21
h10pix90	class											22

Table 6 though Table 16 shows the accuracy achieved by different subsets based on Ig for FS.

Table 6: Accuracy achieved for subset 1 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.7025641	0.7185714	0.7320186	0.7056277
DTC	0.5834336	0.5816619	0.6178125	0.6002115
LSVM	0.645	0.6383	0.6366	0.63875

Table 7: Accuracy achieved for subset 2 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.7623932	0.7457143	0.7436195	0.7164502
DTC	0.5949903	0.6613826	0.6360194	0.6402348
LSVM	0.65	0.65	0.6533	0.65125

Table 8: Accuracy achieved for subset 3 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.8358974	0.8642857	0.8642691	0.8441558
DTC	0.7599919	0.7757821	0.7769008	0.782318
LSVM	0.759	0.7475	0.7646	0.76375

Table 9: Accuracy achieved for subset 4 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.8324786	0.88	0.8770302	0.8603896
DTC	0.7816642	0.7827526	0.8042387	0.7966459
LSVM	0.88	0.7475	0.7726	0.76875

Table 10: Accuracy achieved for subset 5 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	**60-40**	**75-25**	**80-20**
RFC	0.9230769	0.9314286	0.9315545	0.9350649
DTC	0.8846188	0.8644302	0.8775675	0.8781205
LSVM	0.883	0.8775	0.8006	0.88125

Table 11: Accuracy achieved for subset 6 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9111111	0.9285714	0.9303944	0.9318182
DTC	0.8790326	0.8714188	0.8909117	0.8864361
LSVM	0.879	0.881416	0.8873	0.883125

Table 12: Accuracy achieved for subset 7 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9247836	0.94	0.9477958	0.94805192
DTC	0.9109398	0.9144188	0.9215586	0.9162383
LSVM	0.9	0.90416	0.8993	0.898125

Table 13: Accuracy achieved for subset 8 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9418803	0.9471429	0.9535963	0.9577922
DTC	0.9183449	0.9305695	0.9275587	0.9272972
LSVM	0.904	0.905	0.90266	0.90125

Table 14: Accuracy achieved for subset 9 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.948803	0.9428571	0.950116	0.9577922
DTC	0.9070137	0.9305695	0.9275587	0.909999
LSVM	0.906	0.90416	0.90933	0.905

Table 15: Accuracy achieved for subset 10 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9316239	0.9371429	0.9466357	0.948051
DTC	0.899623	0.9022381	0.9086562	0.910017
LSVM	0.903	0.8975	0.91	0.89875

Table 16: Accuracy achieved for subset 11 of Table 5

Technique used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9418803	0.9485714	0.9524362	0.9534632
DTC	0.9070301	0.919239	0.9193146	0.9145792
LSVM	0.914	0.9125	0.92133	0.9125

Table 17 though Table 27 shows the accuracy achieved by different subsets based on Ig for BE.

Table 17: Accuracy achieved for subset 12 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9333333	0.9485714	0.950116	0.9534632
DTC	0.861312	0.8725084	0.8633209	0.876796
LSVM	0.913	0.9116	0.918	0.9125

Table 18: Accuracy achieved for subset 13 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9401709	0.9471429	0.9524362	0.952381
DTC	0.875379	0.8750444	0.8666192	0.8758585
LSVM	0.912	0.9083	0.9086	0.9125

Table 19: Accuracy achieved for subset 14 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.926957	0.9385714	0.9443155	0.9426407
DTC	0.8729629	0.870211	0.8708629	0.8799948
LSVM	0.911	0.9016	0.9073	0.90375

Table 20: Accuracy achieved for subset 15 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9264957	0.9185714	0.9315545	0.9264069
DTC	0.8734058	0.8685641	0.8746661	0.8733429
LSVM	0.907	0.8983	0.9	0.8975

Table 21: Accuracy achieved for subset 16 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9230769	0.92	0.9280742	0.9307359
DTC	0.9050363	0.8974384	0.9022375	0.8937307
LSVM	0.907	0.9	0.8986	0.89875

Table 22: Accuracy achieved for subset 17 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9264957	0.92	0.9280742	0.9253247
DTC	0.9052631	0.8977675	0.8980082	0.905001
LSVM	0.906	0.900083	0.9006	0.899375

Table 23: Accuracy achieved for subset 18 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.9196581	0.9171429	0.9234339	0.9231602
DTC	0.8963239	0.9035607	0.893986	0.901651
LSVM	0.904	0.89583	0.8973	0.89875

Table 24: Accuracy achieved for subset 19 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.876923	0.874286	0.87471	0.875541
DTC	0.900671	0.89117	0.896664	0.900009
LSVM	0.9	0.885	0.8913	0.888125

Table 25: Accuracy achieved for subset 20 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.876923	0.882857	0.87935	0.875541
DTC	0.919351	0.918921	0.914439	0.918536
LSVM	0.9	0.88416	0.8906	0.8875

Table 26: Accuracy achieved for subset 21 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.880392	0.884286	0.844316	0.880952
DTC	0.918974	0.918355	0.90666	0.912073
LSVM	0.889	0.88083	0.884	0.88375

Table 27: Accuracy achieved for subset 22 of Table 5

Techniques used for finding accuracy	Percentage Split			
	50-50	60-40	75-25	80-20
RFC	0.880392	0.885714	0.881671	0.8866831
DTC	0.911635	0.910221	0.914671	0.910403
LSVM	0.888	0.887816	0.8846	0.88375

V. CONCLUSION

In this project work we have applied three machine learning algorithms such as RFC, DTC and Linear SVM based on the Ig value of each attributes of Dengue disease. Based on Ig value we have applied FSM's such as FS and BE to get different subsets of attributes. For each subset we find the prediction accuracy of the three techniques using percentage split as test option. Based on the prediction accuracy achieved by different methods we have identified that:

- The percentage accuracy achieved by DTC for full set of attributes without using FSM's is 92% and the maximum percentage accuracy achieved after using FSM's by subset 8 is 92.73%.
- The percentage accuracy achieved by RFC for full set of attributes without using FSM's is 95.45% and the maximum percentage accuracy achieved after using FSM's by subset 8 is 95.78%.
- The percentage accuracy achieved by Linear SVM for full set of attributes without using FSM's is 92.6% and the maximum percentage accuracy achieved after using FSM's by subset 8 is 92.13%.
- Important attributes obtained after performing FSM's on all the three techniques are trees, Xmin, Xmax, temp, temp90, Ymin, Ymax and humid90.

REFERENCES

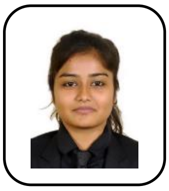
1. Hales, et al. "Potential effect of population and climate changes on global distribution of dengue fever: an empirical model." The Lancet 360.9336 , 2002, 830-834.
2. McCarthy, et al. "Climate change 2001: impacts, adaptation, and vulnerability" Intergovernmental Panel on Climate Change, Vol. 2, Cambridge University Press, 2001.
3. Van Lieshout, et al. "Climate change and malaria: analysis of the SRES climate and socio-economic scenarios." Global Environmental Change , 14.1 ,2004, 87-99.
4. R. Sanjudevi and D. Savitha, "Dengue fever prediction using classification techniques" International Research Journal of Engineering and Technology (IRJET) , Volume: 06 ,Issue: 02 ,Feb 2019,Page-558-563.
5. N. Rajathi, et al., "Early Detection of Dengue Using Machine Learning Algorithms." International Journal of Pure and Applied Mathematics , Volume :118, Issue 18, 2018, 3881-3887.
6. Tate, A., et al. "Prediction of dengue diabetes and swine flu using random forest classification algorithm." Int. RJ Engg. Tech ,4 ,2017, 685-690.
7. Yusof, et al. "Dengue outbreak prediction: A least squares support vector machines approach." International Journal of Computer Theory and Engineering ,3.4 ,2011,485- 489.
8. Yusof, et al. "Dengue outbreak prediction: A least squares support vector machines approach." International Journal of Computer Theory and Engineering ,3.4 , 2011, 485- 489.
9. Guyon, Isabelle, and Andre Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.
10. P. Manivannan and P. Isakki Devi, "Dengue fever prediction using K-means clustering algorithm" , 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS),Pages:1-5,Year:2017
11. N. M. Tahir, A. Hussain, S. A. Samad, K. A. Ishak and R. A. Halim, "Feature Selection for Classification Using Decision Tree," 2006 4th Student Conference on Research and Development, Selangor, 2006, Pages: 99-102.

AUTHORS PROFILE

Dr. Raghavendra S. is currently working as Associate Professor in the Department of Computer Science and Engineering at CHRIST DEEMED TO BE UNIVERSITY, Bangalore. He completed his Ph.D. degree in Computer Science and Engineering from VTU, Belgaum, India in 2017 and has 15 years of teaching experience. His interests include Data Mining and Big data.

Subhram Dasgupta completed his B. Tech in the Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore in 2019**.** His interests include Data Mining and Big data.

Naman Sharma completed his B. Tech in the Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore in 2019**.** His interests include Data Mining and Big data.



Sweta Sinha completed her B. Tech in the Department of Computer Science and Engineering from CHRIST DEEMED TO BE UNIVERSITY, Bangalore **in** 2019**.** His interests include Data Mining and Big data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)