International Journal of Science, Engineering and Management (IJSEM)
Vol 2, Issue 12, December 2017

A Survey on Prediction of Dengue Fever Using Data Mining Techniques

[1] Ms.S.Freeda Jebamalar, [2] Dr.A.Anitha
[1]PG Scholar, [2] Professor
[1][2] Department of Information Technology, Francis Xavier Engg.College, Tirunelveli.

***Abstract:--* Dengue is one of the most life threatening disease in Tamil Nadu.This disease affects life of many people in the state.However government taken many precaution to control this disease.However It cannot be controlled fully by government.The main cause for this disease is female mosquitoes.It is typically found in widespread hot regions.The symptoms for this disease will vary from one person to another person.Dengue infection has endangered 2.5 billion and more population all over the world[3].The major symptoms for dengue are intermitten fever, headache,joint pain,bleeding,vomiting,etc.,The diagnosis of disease is a vital and intricate job in medicine.Human however intelligent they may be,they are not experts on their own.In this paper ,a survey is made on application of data mining techniques like Naive bayes,J48 classifier and clustering algorithm like K-mediods,Dbscan ,k-means etc.,accurately predicting dengue disease .It also reduces time for prediction of disease.and reduces time for prediction of disease.**

***Keywords:--* Prediction,Naïve bayes,J48(Decision tree),K-means,K-mediods,Dbscan.**

1. INTRODUCTION

Data mining is a process of identifying novel, potentially useful, valid and ultimately understandable patterns in data. The data has many applications in the field of telecommunication industry, biological data analysis and much more .It makes use of Artificial Intelligence, machine learning and database management to extract new patterns from large dataset and the knowledge associated with these patterns. This data mining can be used in the scientific side to detect patterns from medical datasets. The abundant case sheets and patients records stored in the hospitals have vital information like symptoms and vitals of the patients. These records can be used to identify the patterns. Data mining utilize the history of records to predict the disease with high accuracy. Dengue fever is one of the most vulnerable diseases which affects life of many people .It is caused by Ades Aegypti female mosquitoes. It has symptoms like headache, intermittent fever, bleeding, nausea, vomiting, joint pain, etc.,It leads to the minimization of blood count which would leads to death. With the help of many classifiers like Naïve Bayes ,J48 Rep Tree, SMO, Neural Network, K Nearest Neighbors(KNN),classify the dataset. After the classification some performance evaluation is done using measures like accuracy, precision, sensitivity, specificity, area under ROC ,F-measure, cost-Benefit analysis .

2.LITERATURE SURVEY:

2.1. Analysis using Informatica Tools(2016)[1]:

This paper refers the dengue infection occurred in the Dharmapuri district.[1].The author of this paper took the dataset from Dharmapuri health center and used Informatica tool for the analysis and to detect the affected patients earlier. For better accuracy this tool was used to differentiate the dengue affected person and healthy person. Informatica tool used ETL which would convert heterogeneous data into homogenous.[2].In the first phase of extraction data would be stored in the temporary storage area where the duplication is removed with the help of validation rules. In the second phase transformation is to be done. This phase include data formatting, resorting rows& columns, splitting the data. The main purpose of transformation phase is to make data in uniform manner. In the last phase, load the data into database. Informatica tool is mainly used for the purpose of big data analytics. While others are not. It has the ability to arranged the session into worklets and workflows .Job Monitoring and recovery capability would help to identify the slowest running jobs and recover the job by restarting from the failure step. The source file is imported into csv format. The data is subjected to preprocess in the router transformation because it would help to store unprocessed data for future use. Then providing one constraint which drops the unnecessary data.70% were found out to be healthy of 100% samples. The negative results are stored in the flat-file database and positive results are stored in oracle database. This paper finally concluded that the children below 10 are prone to be affected as compared to adults, with prediction accuracy of about 70%.

2.2 Decision Tree and Support vector Machine (March 2017)[3]:

In this paper the algorithmssuch as Naive Bayes,J48,SMO,REP Tree and Random tree are employed for the purpose of getting a better accuracy in dengue prediction.With the explorer interface and knowledge flow interface in data mining .Among the above algorithms Naïve Bayes and J48 achieved the maximum accuracy.The attributes are id,fever,bleeding,flu,fatigue,etc.,.Four types of precision were employed in it. TN-case was negative tuples predicted as negative.TP- case was positive tuples predicted as positive.FN- case was positive tuples predicted as negative.FP-case was negative tuples predicted as positive.The dataset was saved as text file then it was imported into the Excel spreadsheet.Missing values were replaced as during preprocessing.FisherFiltering techniques are used to rank the input attributes according to their relevance.After the computation of Fisher score,attribute having largest score among the all other attributes is selected .

2.3 Decision Tree and k-means clustering (2014) [3]:

In this paper, the dataset were collected from the two hill areas in east Godavari which is located in the district of Andhra Pradesh. The attributes were based on the education, income, hereditary factors, area location, drainage facilities, drinking water facilities, toilet facilities, waste disposal, electricity, approaches to hospitals, roads, educational institutional, livelihood etc., Applying unsupervised learning(K-means) approach to cluster the similar dataset for the better understanding of data. In this paper clustering is done based on hereditary factors. From the above, dataset was built only with the records of people who tends to become insolvent. The above problem can be solved by classification technique(Decision tree) and achieved the accuracy level of 97% .It implies 97% of correctly classified instances.

2.4 NEW INTELLIGENT BASED APPROACH FOR COMPUTER-AIDED DIAGNOSIS[IEEE][Jan 2012][4]:

In this paper, the probable cases of the DF(Dengue Fever) are identified by the clinical test such as ELISA(**e**nzyme-**l**inked **i**mmuno**s**orbont **a**ssays).The dataset was obtained from the central and western states of India. The clinical dataset were grouped into four catagories.DS1, DS2, DS3, DS4.In the DS1, they took 646 adults (greater than or equal to 16 yrs)records. The authors considered the clinical symptoms and the lab features for the dataset DS1.Out of 646 cases, 256 patients were affected by Dengue fever and rest of were not affected. In DS2 the clinical symptoms were considered and ignoring the lab features. In DS3 dataset includes 398 children (5 to 15 yrs).The authors consider the clinical symptoms and the lab features. About 93 children were affected by dengue. DS4 was a part of DS3, considered only the clinical symptoms. In the above for dataset the NM method was applied and achieved the accuracy of 100% in the DS1and DS3.NM is the Universal tool for detecting the effectness of disease.NM tool has generated decision tree with the accuracy of 100% in children and adult using both the clinical and laboratory features.The performance metrics were AUC(area under ROC),SE(sensitivity),SP(specificity).Arthalgia was found to be the influencing factor in children and adult

2.5 DENGUE INCIDENCE USING CLUSTERING(k-MEANS) BASED REGRESSION ON CLIMATE DATA[November IEEE 2016] [5]:

These paper refered the dengue infected cases in Malaysia has significant increase during 2011-'15.It includes 1,20,000 cases and most of them were reported from the Selangor. The data obtained from the Ministry of Health(MOH),Malayasia.All the data are uploaded in to the Maria DB database. This paper preferred the climate factors for prediction of dengue. It includes the environmental factors such as relative humidity, rainfall, temperature. Relative humidity is one of the important factor strongly influences (70%-80%) the survival of the mosquitoes. Another reason is the optimum rainfall lead to the good source for the warm ambient temperature for the mosquito's gonotrophic cycle. First step was cluster the data. and then the data was normalized with the value 0 and 1.Then the data is partitioned into the value of K=3 by K-means algorithm in order to reduce the error. There many types in value K. Here average silhouette width method used in this paper. Regression technique used to produce the high accuracy of 92% in the dengue outbreak.

3.SUMMARY OF LITERATURE SURVEY:

SNO	TECHNIQUES USED	DATASET	ATTRIBUTES	PREDICTION METRICS	ACCURACY
1	Informatica Techniques (number Of tranfor	Dharmapuri Health center	1. Headache 2. Body pain 3. Abdominal pain 4. continuous vomiting 5. fever 6. Bleeding tendency	7 symptoms with the temperature	70%

	mation)		7. reduction in WBC platelets	of the fever.	
2	Naïve Bayes,J48(Decision tree),Rep Tree	Not specified	1. id 2. Fever 3. bleeding 4. flu 5. fatigue	Fisher Filtering	**92%**
3	k-Means clustering and Decision Tree classification	Two Hills areas in East Godavari (A.P)	1. Income 2. Employment 3. EnvironmentalFactor 4.Members in the Family 5.sanitation 6.Education 7.hereditary 8.Drinking waterFacility 9.Age	Environmental condition	**97%**
4	**NM**Tool generated Decision Tree	Not specified	1.Clinical symptoms 2.laboratory Features 3.ELISA test	AUC(Area under ROC curve), SE(sensitivity) ,SP(specifivity).	**100%**
5	K-Means clustering	Ministry of Health, Malayasia	1. Rainfall 2. Temperature 3. Relative Humidity	Temperature	**92%**

4. CONCLUSION AND FUTURE ENHANCEMENT:

From the literature survey, it has been concluded that data mining techniques aremore effective in predicting the dengue disease with accuracy ranging from 70% to 100% .In future,it is proposed to improve the accuracy by incorporating some more features and latest datamining techniques.

REFERENCES:

[1]. ”S.Stanly Leena Princy,A.Muruganadam”,”An Implementation of Denguen Fever Disease spreadusing Informatica Tool with Special Reference to Dharmapuri District(IJIRCEE) “,Vol 4,Issue 9,Sep 2016.pg 16215-16222.

[2]. “Dr.Arun Kumar P.M,Associate Professor,Chitra Devi .B,Karthick .P,Ganesan M,Madhan A.S,”Dengue Disease Prediction Using Decision Tree and Support Vector Machine”,(ICET’17)-Special Issue-March 2017,pp_60-63.

[3].”N K Kameswara Rao,Dr.G P Saradhi Varma,Dr.M .Nagabhushana Rao, Assoc Professor IT Dept,SRKR Engineering college,Bhimavaram ,AP,” Classification Rules Using Decision Tree for Dengue Disease”,vol 3,Issue 3,(IJRCCT),March 2014,pp_340-343.

[4].”Shermon S.Mathumuthu,Vijanth S.Asirvadam,Sarat c.Dass,Balvin S.G,Losini T”,”Predicting Dengue Incidences using cluster Based Regression on climate Data”,

[5].”Vadrevu Sree Hari Rao,Senior Member IEEE and Mallenahalli Naresh Kumar”,”A New Intelligence Based Approach for Computer-Aided Diagnosis of Dengue Fever”,Vol 16,No 1,Jan 2012,IEEE TRANSACTION ON INFORMATION TECHNOLOGY IN BIOMEDICINE.