COMPARISION AND COMBINATION OF MINING TECHNIQUES FOR GENE ANALYSIS TO IDENTIFY DENGUE

Dave Kaveri Atulbhai

M.E. Student, Computer Engineering, Gujarat Technical University, Gujarat, India

ABSTRACT

Data Mining is focused on deriving knowledge from multiple data formats using intelligent analytics techniques. Data mining is a field having various techniques that converts the raw data into useful information in various research fields. There are two primary goals for data mining prediction and description. Prediction involves fields or variables in the data sets to predict unknown or future values of other diseases possibilities. On the other hand description involves finding of pattern describing the data that can be present in knowledge base provided for disease prediction. We can predict diseases like hepatitis, Lung cancer liver disorder, breast cancer or heart diseases, diabetes etc, it helps in finding the patterns to decide future trends in medical field. We can use in finding the knowledge and prediction, detection of Dengue as well.

Keywords :-*Data Mining, Raw Data, Information, Knowledge Discovery, Prediction, Description, Medical Analysis, Dengue.*

1. INTRODUCTION:

Information technology has generated large amount of data-base and huge amount of data in various research fields. To research in knowledge mining has given rise to store data and manipulate previously stored data for further decision making process. Dengue is a threatening disease caused by female mosquitoes. It is typically found in widespread hot regions. From long periods of time, Experts are trying to find out some of features on Dengue disease so that they can rightly categorize patients because different patients require different types of treatments [2]. There are various data mining techniques available with suitable dependent on domain application. By using data mining we can examine large amount of routine samples collected in disease prediction. Best results are achieved by balancing knowledge of experts for describing the problem and goals with search capabilities. Hospitals must also want to minimize cost of clinical test. It can be achieved by employing appropriate computer based information and decision sport system. Here, data mining plays an important role to give many results faster and accurate by using various algorithms [8].

2. DATA MINING PROCESS AND MEDICAL SCIENCE:

Data mining is the process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [8].

2.1 Data mining involves the following Steps:

1. Problem definition: The first step is to identify goals.

2. Data exploration: All data needs to be consolidated so that it can be treated consistently.
3. Data preparation: The purpose of this step is to clean and transform the data for more robust analysis.
4. Modeling: Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
5. Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

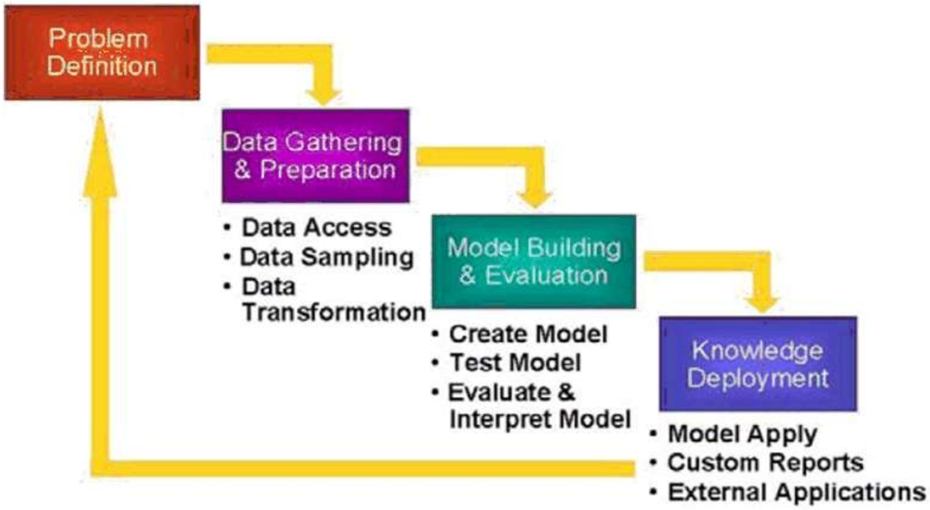


Fig -1: Data Mining Process

2.2 Data mining for Medical Science:

Data mining is a relatively new field of research whose major objective is discovering knowledge from large amounts of data. In medical and health care areas a large amount of data is becoming available due to regulations and the availability of computers and technology. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. A major objective is to evaluate data mining tools in medical and health care applications to develop a tool that can help make timely and accurate decisions.

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acute care hospital may generate five terabytes of data a year [7]. The ability to use these data to extract useful information for quality healthcare is crucial. Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place.

The application of artificial intelligence in healthcare is relatively new. Data mining can be applied to the medical databases, which will predict or classify the data with a reasonable accuracy. For a good prediction or classification the learning algorithms must be provided with a good training set from which patterns and rules can be extracted to help classify the testing dataset. We can predict diseases like hepatitis, Lung cancer liver disorder, breast cancer or heart diseases, diabetes, arbovirus disease etc, it helps in finding the patterns to decide future trends in medical field.

3.PREDICTION OF AN AIRBOVIRUS – DENGUE DISEASE:

Dengue infection is an epidemic disease typically found in tropical region. Symptoms of the disease show rapid and violent for patients in a short time. The World Health Organization (WHO) classifies the dengue infection as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF). Symptoms of DHF are divided into 4 types. The problem might be happen when an expert misdiagnoses dengue infection. For Example, an expert diagnosed a patient as non dengue or DF even if a patient was a DHF patient. That might be the cause of dead if patient did not receive treatment. Therefore, we selected data mining approach to solve this problem [2].

For properly categorizing our dataset, different classification techniques are used. These techniques are Naïve Bayesian, REP Tree, Random tree, J48, SMO, SVM, Decision Tree Approach, and Spatial Data Analysis as well.

In this survey study, [1] Represents Comparative Analysis of Machine learning techniques for classification of Arbovirus - Research efforts have reported with increasing confirmation that the support vector machines (SVM) have greater accurate diagnosis ability. The accuracy of svm outperforms the Naïve bayes technique. As shown in the experiment results, support vector machine has the highest classification precision most of the time. However support vector machine is very time consuming because of more parameters, demands more computation time. [2] Represents Data Mining of Dengue Infection Using Decision Tree in which each dataset consists of more than 400 attributes. To accomplish the knowledge discovery task, we consider employing decision tree as a data mining tool. We propose a set of meaningful attributes from the temporal data. Our experiments are divided into 4 parts. The first two experimental results show the useful knowledge to classify dengue infection from 2 different datasets respectively. Another objective of this research is to detect the day of defervescence of fever which is called day0. At the end we obtained very low accuracy in day-4 as we found that the tree is over fit. The experimental results shown that the decision tree approach did not suit this task thus we think we should to select a new classification approach in the future works. [3] Represents the classification techniques to determine the population of Dengue fever infected cases in Jhelum district and in surrounding areas geographically. So, we can compare performance of different classification techniques. Objective of this study also includes the comparison of different classification algorithms with the help of graphs, based on our dataset. We have implemented all the techniques by using weka tool and all the procedure of implementation is within it. At the end, After analysis of our dataset with each technique we are paralleling them in the conclusion. When we have done the comparison among all of them we concluded that naïve Bayes Technique is greatest among all others. As the accuracy of Naive bayes is 92% which was biggest of all. Naïve Bayes is the best also for the aim that it gives the probability and efficiency while Random Tree and REP Tree do not give us probability. It has concluded that NB and J48 are the top performance classifier techniques by way that, they has achieved an accuracy of 92% and 88%, takes fewer time to run and shows ROC area=0.815, and had smallest error rate. We can find the following comparison chart below.[4] Represents implementation the automated system to reorganize the Dengue fever using the Microscope blood image report. Microscope image report as input and signals are filtered and the feature characteristics are extracted, the features are fed to the neural networks. Classification is carried using and the Back Propagation Network (BPN).Which gives the 98% accurate result with a short period. [5] Represents prediction of chronic kidney disease. Study is focused on the usage of classification techniques in the field of medical science and bioinformatics. The main focus of this paper is Chronic-Kidney-Disease prediction using weka data mining tool and its usage for classification in the field of medical bioinformatics. Algorithms have been compared with classification accuracy to each other on the basis of correctly classified instances, time taken to build model, time taken to test the model, mean absolute error, K appa statistics and ROC Area. In the experiments Multilayer perception algorithm gives better classification accuracy and prediction performance to predict chronic kidney disease (CKD) using relevant dataset available at UCI machine learning repository. [6] Represents a study of different data mining techniques that can be employed in automated HCV infection prediction systems. The system extracts hidden knowledge from a historical HCV patient's database. The models are trained and validated against a test dataset. Three classification techniques which are Naïve Bayes, decision trees and neural networks applied on three HCV database of different size, then determine Performance accuracy and effectiveness of each technique using Lift Chart. Performance analysis of these classifiers is summarized in Table.

4. METHODOLOGY, EXPECTED OUTCOMES AND IMPLEMENTATION STRATEGIES:

Data mining is a relatively new field of research whose major objective is discovering knowledge from large amounts of data. In medical and health care areas a large amount of data is becoming available due to regulations and the availability of computers and technology. On the one hand, practitioners are expected to use all this data in their work but, at the same time, such a large amount of data cannot be processed by humans in a short time to make diagnosis, prognosis and treatment schedules. A major objective is to evaluate data mining tools in medical and health care applications to develop a tool that can help make timely and accurate decisions [10].

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been estimated that an acue care hospital may generate five terabytes of data a year [10]. The ability to use these data to extract useful information for quality healthcare is crucial. Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place.

The application of artificial intelligence in healthcare is relatively new. Data mining can be applied to the medical databases, which will predict or classify the data with a reasonable accuracy. For a good prediction or classification the learning algorithms must be provided with a good training set from which patterns and rules can be extracted to help classify the testing dataset. We can predict diseases like hepatitis, Lung cancer liver disorder, breast cancer or heart diseases, diabetes, arbovirus disease etc, it helps in finding the patterns to decide future trends in medical field [9].

4.1 Work Flow with Methodology:

Here, we described proposed work with expected outcome. After pre-processing the data sets apply SVM algorithm and get results then we will apply Naïve Bayes to the same data sets, make model, labeled out gene with respective expressions and get resulted gene data which are responsible for dengue.

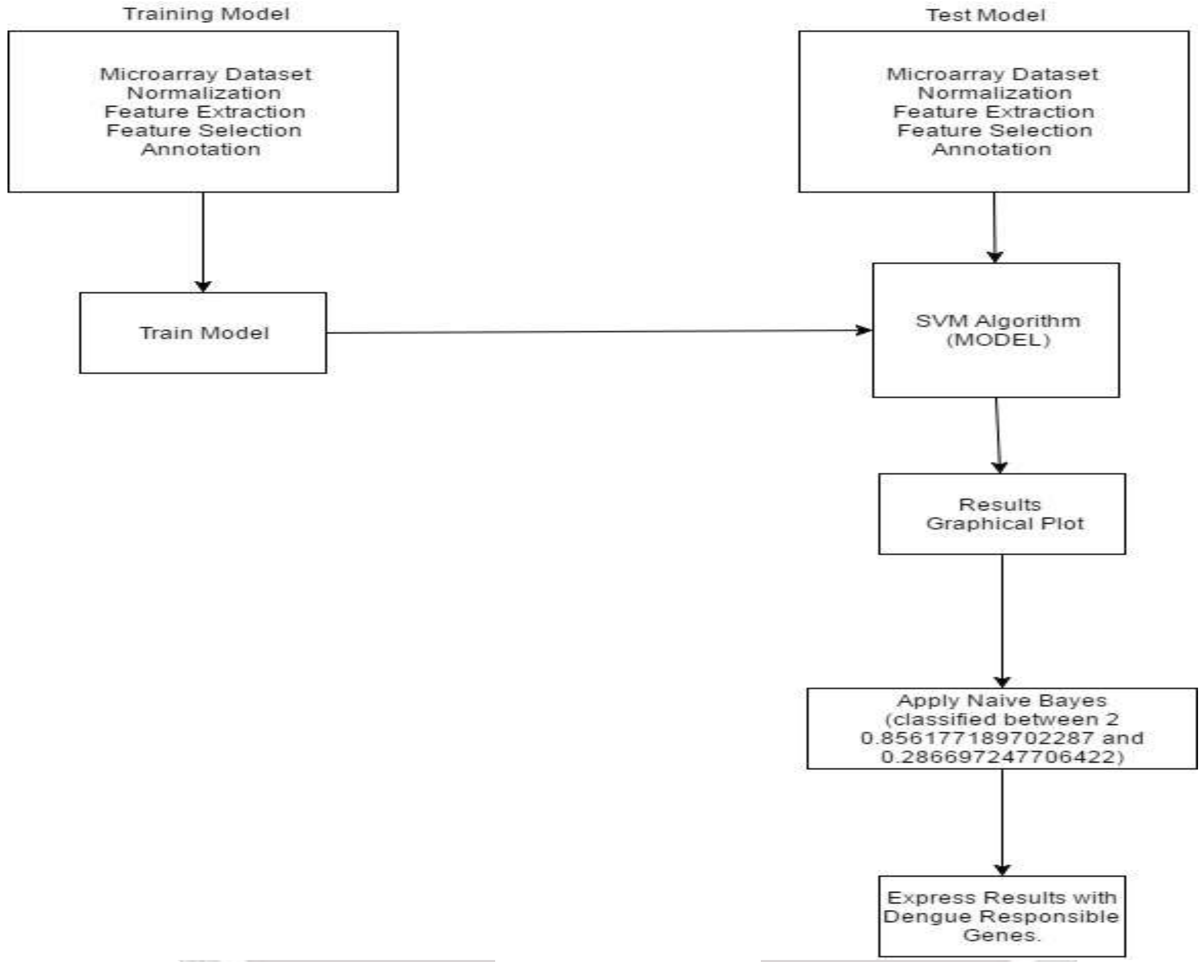


Fig – 4.1 Flow Diagram

4.2 Expected Outcome:

Classified highly down regulated gene which are responsible for dengue**.**

4.3 Implementation Strategies:

Input Dataset	Reports of Patients
Total Dataset	758 MB to 1256 MB
Software	R Studio
Library	R Library
Programming Language	R Language by R Studio

4.4 IMPLEMENTATION AND RESULT:

Let’s consider following work flow Algorithm for mentioned system:

1. Start with Micro Array data set.

#Plot Histogram and Boxplot before preprocessing.

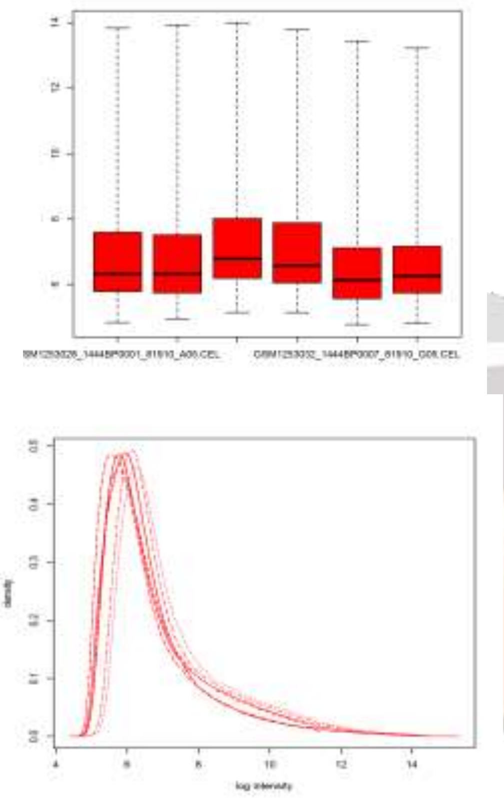


Fig – 4.1 (A): Before Normalization

2. Apply Normalization with RMA.

3. Express data in log2 expression as data not in a feasible range.
   #Plot Histogram and Boxplot after Normalization.

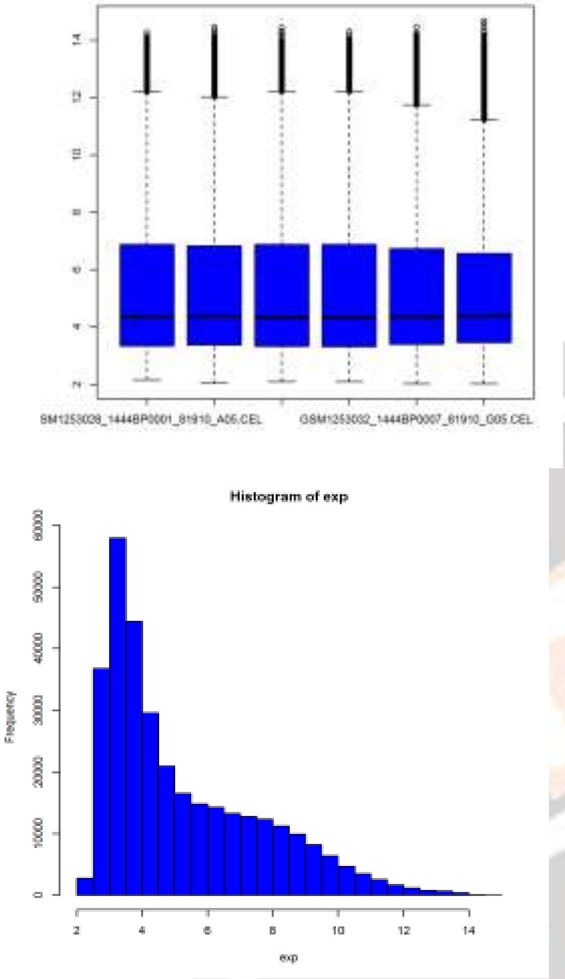
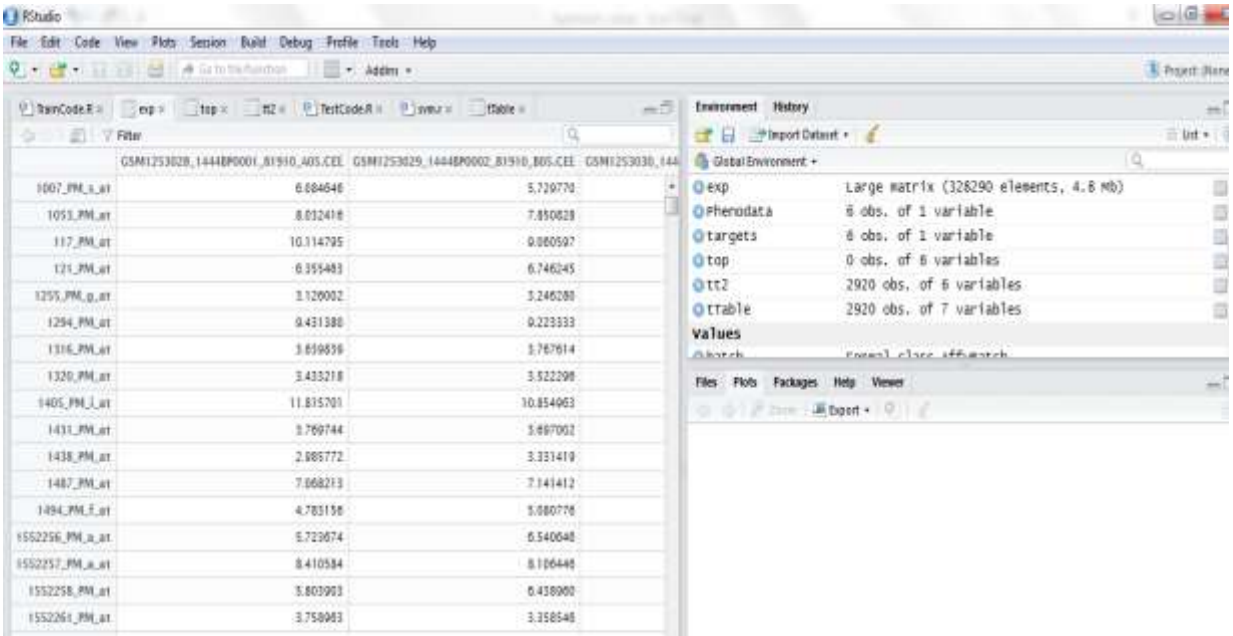


Fig – 4.1 (B): After Normalization

4. Feature Extraction.
   #Apply eBayes Algorithm for p.value, Log FC value, Adj p.value calculations.



	GSM1253028_1444BP0001_81910_A05.CEL	GSM1253029_1444BP0002_81910_B05.CEL	GSM1253030_144
1007_PM_s_at	6.684646	5.729770	
1053_PM_at	8.032418	7.850828	
117_PM_at	10.114795	9.060597	
121_PM_at	6.355483	6.746245	
1255_PM_g_at	3.126002	3.246280	
1294_PM_at	9.431380	9.223333	
1316_PM_at	3.659839	3.767614	
1320_PM_at	3.433218	3.522298	
1405_PM_i_at	11.835701	10.854063	
1431_PM_at	3.769744	3.697002	
1438_PM_at	2.985772	3.331419	
1487_PM_at	7.068213	7.141412	
1494_PM_f_at	4.783156	5.080776	
1552256_PM_a_at	6.723674	6.540648	
1552257_PM_a_at	8.410584	8.106446	
1552258_PM_at	5.803903	6.438960	
1552261_PM_at	3.758963	3.358548	

exp	Large matrix (328290 elements, 4.8 Mb)
Phenodata	6 obs. of 1 variable
targets	6 obs. of 1 variable
top	0 obs. of 6 variables
tt2	2920 obs. of 6 variables
tTable	2920 obs. of 7 variables
Values	
batch	Formal class AffyBatch

Fig – 4.1 (C): Feature Extraction

5. Feature Selection.
   #values with p.value < 0 used to select differently expressed or say significant genes.
   #Log FC calculated bay p.value analysis.

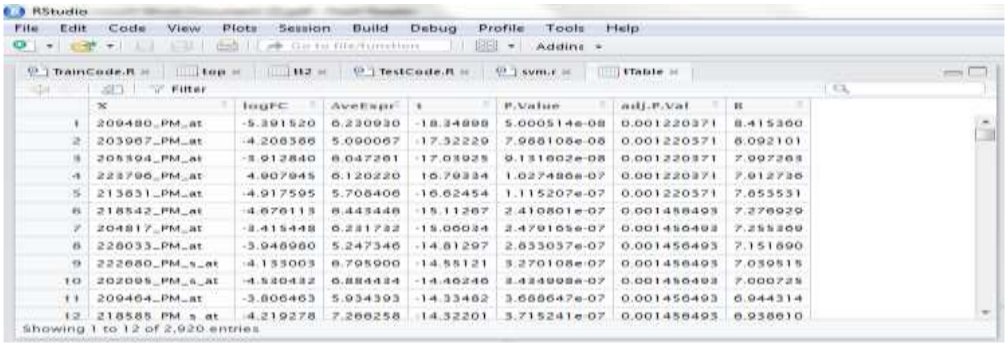
```
If ( logFC < 0)
Then negative logFC value which represents logarithmic foldness of down regulation
Else if ( logFC > 0)
Then positive logFC value which represents logarithmic foldness of up regulation
```

Note: LogFC = 0 consider as default bu LIMMA Package Library.

#Genes which shows negative logFC value denotes the downregulation of their corresponding gene function

#which means these genes are infected during disease.

#This table shows the predicted negative logFC values

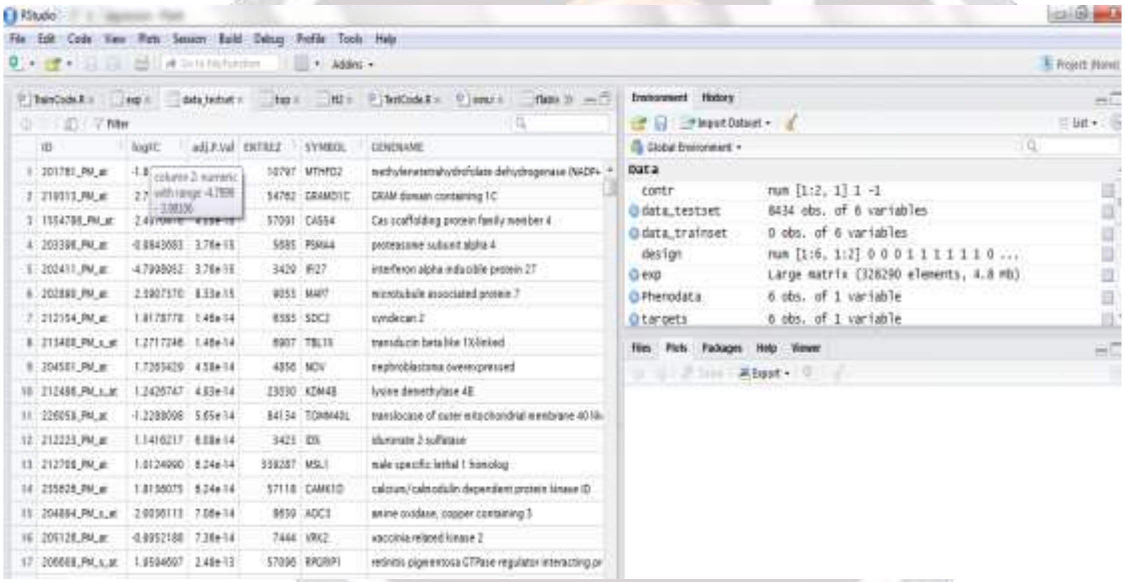


	X	logFC	AveExpr	t	P.Value	adj.P.Val	B
1	209480_PM_at	-5.391520	6.230930	-18.34898	5.000514e-08	0.001220371	8.415360
2	203967_PM_at	-4.208366	5.090067	-17.32229	7.968108e-08	0.001220371	8.092101
3	205394_PM_at	-3.912840	6.047261	-17.03925	9.131602e-08	0.001220371	7.997263
4	223796_PM_at	4.907945	6.120220	16.79334	1.027486e-07	0.001220371	7.912736
5	213831_PM_at	-4.917595	5.708406	-16.62454	1.115207e-07	0.001220371	7.855531
6	218542_PM_at	-4.676113	6.443446	-15.11267	2.410801e-07	0.001456493	7.276929
7	204817_PM_at	-3.415448	6.231732	-15.06034	2.479165e-07	0.001456493	7.255309
8	228033_PM_at	-3.948980	5.247346	-14.81297	2.833037e-07	0.001456493	7.151690
9	222680_PM_s_at	-4.135003	6.795900	-14.55121	3.270108e-07	0.001456493	7.039515
10	202095_PM_s_at	-4.520432	6.884434	-14.46246	3.434908e-07	0.001456493	7.000725
11	209464_PM_at	-3.806463	5.934393	-14.33462	3.688647e-07	0.001456493	6.944314
12	218585_PM_s_at	-4.219278	7.266258	-14.32201	3.715241e-07	0.001456493	6.938610

Showing 1 to 12 of 2,920 entries

Fig – 4.1 (D): Feature Selection

6. Apply Annotation.
   #generated data set having Database & CDF (Chip Description Format), we consider Database as Database having gene name and description as well.
   #to convert in readable format we are using Annotation.



	ID	logFC	adj.P.Val	ENTREZ	SYMBOL	GENENAME
1	201761_PM_at	-1.8	[illegible]	10797	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+
2	218313_PM_at	2.7	[illegible]	54762	GRAMD1C	GRAM domain containing 1C
3	1554788_PM_at	2.49	[illegible]	57091	CASS4	Cas scaffolding protein family member 4
4	203396_PM_at	-0.8843683	3.76e-15	5685	PSMA4	proteasome subunit alpha 4
5	202411_PM_at	-4.7998062	3.76e-15	3429	IFI27	interferon alpha inducible protein 27
6	202890_PM_at	2.5907570	8.33e-15	9053	MAP7	microtubule associated protein 7
7	212154_PM_at	1.8178778	1.46e-14	6385	SDC2	syndecan 2
8	213400_PM_s_at	1.2717246	1.46e-14	6907	TBL1X	transducin beta like 1X-linked
9	204501_PM_at	1.7265429	4.58e-14	4856	NOV	nephroblastoma overexpressed
10	212496_PM_s_at	1.2426747	4.83e-14	23030	KDM4B	lysine demethylase 4B
11	226058_PM_at	-1.2288098	5.65e-14	84134	TOMM40L	translocase of outer mitochondrial membrane 40 lik
12	212223_PM_at	1.1416217	6.88e-14	3423	IDS	iduronate 2-sulfatase
13	212708_PM_at	1.0124990	8.24e-14	339287	MSL1	male specific lethal 1 homolog
14	235828_PM_at	1.8158075	8.24e-14	57118	CAMK1D	calcium/calmodulin dependent protein kinase ID
15	204894_PM_s_at	2.0036113	7.08e-14	8639	AOC3	amine oxidase, copper containing 3
16	205126_PM_at	-0.8952188	7.38e-14	7444	VRK2	vaccinia related kinase 2
17	206668_PM_s_at	1.9594697	2.48e-13	57096	RPGRIP1	retinitis pigmentosa GTPase regulator interacting pr

column 2: numeric with range -4.7898 - 3.08336

```
Data
  contr          num [1:2, 1] 1 -1
  data_testset   8434 obs. of 6 variables
  data_trainset  0 obs. of 6 variables
  design         num [1:6, 1:2] 0 0 0 1 1 1 1 1 1 0 ...
  exp            Large matrix (328290 elements, 4.8 Mb)
  Phenodata      6 obs. of 1 variable
  targets        6 obs. of 1 variable
```

Fig – 4.1 (E): Annotation

7. Check Duplication & Redundancy.

8. Create Train Model.

9. Apply pre-processed test dataset as input to machine learning algorithm.

10. SVM Algorithm with e1071 Package in R.Studio.

11. Analyze Generated results and graphical plot.

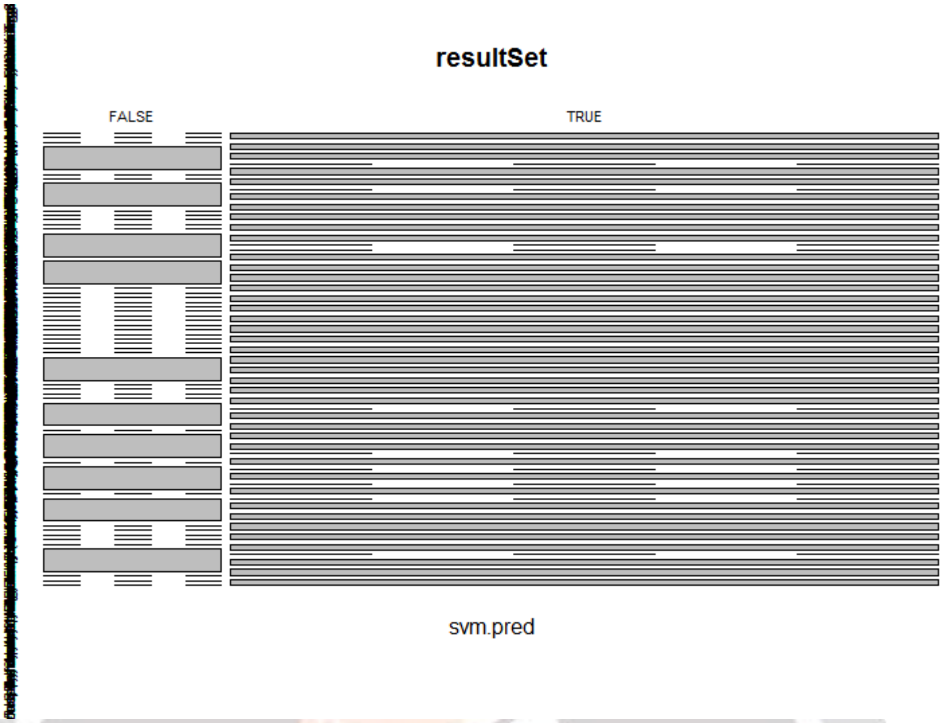


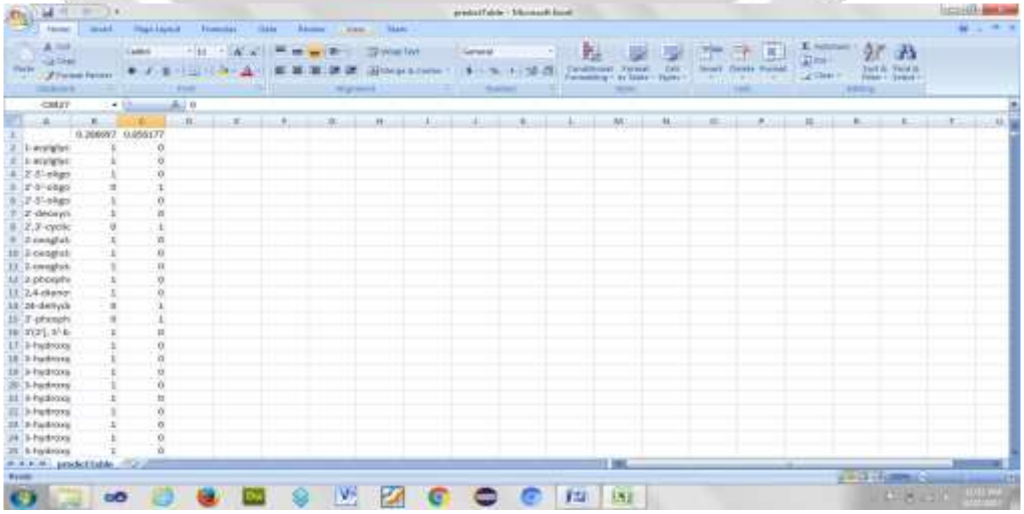
Fig – 4.1 (F): SVM Prediction Result Plot.

12. Interchange the row and column of csv file with header names - Name and Result.

13. Apply Naive Bayes.

14. Result classified between 2 values 0.856177189702287 and 0.286697247706422.

15. Give the value of correctly interpreted down regulated genes.
    #Express Results with Dengue Responsible Genes.



	A	B	C
1		0.286697	0.856177
2	1-acylglyc	1	0
3	1-acylglyc	1	0
4	2',5'-oligo	1	0
5	2'-5'-oligo	0	1
6	2'-5'-oligo	1	0
7	2'-deoxyri	1	0
8	2',3'-cyclic	0	1
9	2-oxoglut	1	0
10	2-oxoglut	1	0
11	2-oxoglut	1	0
12	2-phospho	1	0
13	2,4-dienoy	1	0
14	24-dehydr	0	1
15	3'-phosph	0	1
16	3'(2'), 5'-b	1	0
17	3-hydroxy	1	0
18	3-hydroxy	1	0
19	3-hydroxy	1	0
20	3-hydroxy	1	0
21	3-hydroxy	1	0
22	3-hydroxy	1	0
23	3-hydroxy	1	0
24	3-hydroxy	1	0
25	3-hydroxy	1	0

Fig – 4.1 (G): Prediction Result.

5. COMPARISION & COMBINATION:

The best accuracy is achieved by svm - 0.9042. The good performance of SVM indicates that they can possibly be adjusted to improve specific ad hoc methods for prediction of susceptibility to complex diseases such as Dengue. The total accuracy for 10-fold and 100 cross-fold validation on training data for svm was 90.3 in both cases. We show that sensitivity of naIve Bayes is more than svm in many classification tasks in bioinfonnatics and related scientific fields. The accuracy of svm outperforms the Naive bayes technique. As shown in the experiment results, support vector machine has the highest classification precision most of the time .However support vector machine is very time consuming because of more parameters, demands more computation time. As well we have calculated that SVM having more risk rate compare to naïve bayes & efficiency of naïve bayes is little good than all the methods for bio medical gene analysis.

Measures	***SVM (%)***	***Naïve Bayes (%)***
Sensitivity	47.23	98.77
Efficiency	72.77	94.05
Accuracy	90.18	78.05
Risk Rate	83.43	15.98

Table – 5.1 (A): Comparisons

In this mentioned model we have combined most suitable methods for medical data sets – SVM & Naïve Bayes. As per comparative analysis we can justify that SVM is little more accurate with high risk rate while Naïve Bayes having less accuracy with more efficiency and lowest risk rate. So, we classified gene with more attribute expressions by SVM then apply Naïve Bayes to classified gene in respected range to get dengue infected data from gene data sets. We can increase Efficiency of this model by 29.48% with 70.52% Accuracy.

6. CONCLUSION:

Firstly, there exists a wide class of algorithms and techniques for information extraction and knowledge discovery in medical science. Best results are achieved by balancing knowledge of experts for describing the problem and goals with search capabilities. Hospitals must also want to minimize cost of clinical test. It can be achieved by employing appropriate computer based information and decision sport system. Here, data mining plays an important role to give many results faster and accurate by using various algorithms.By analyzing the different techniques in mentioned researches we can say that more accurate method for medical science is Naïve Bayes as classification methods is more useful in medical science and disease prediction. But, we can combine different methods to get accurate knowledge discovery. Previously studies are along with comparative analysis only.

This paper presents a generic survey of the Dengue Disease Prediction using different data mining techniques. All techniques we explained are widely using in medical science. Data set for dengue prediction is DNA microarray data which have information of gene's expression responsible for dengue virus' successful attack. This data is available on GEO dataset NCBI. For this project .cel format is used as mostly microarray chip machines saved information in this format. Before doing any analysis, Normalization of these data is necessary. Hence, after analyzing all the researches we can conclude that combination of different methods can help to acquire more accurate results.

We apply RMA to normalize microarray data. Then, features selection by using Wrapper Method. We will improve result of detection and diagnose relevant data using 2 way methods – SVM for classification on gene expressions and Naïve Bayes to classified dataset for high level of classification and accurate and efficient results about dengue detection. We can increase Efficiency of this model by 29.48% with 70.52% Accuracy. In future we will predict survivability of patients on basis of gene expressions and their value changes.

4. REFFERENCES:

[1] Shameem A. Fathima and NisarHundewale, Senior Member, “Comparative Analysis of Machine learning Techniques for classification of Arbovirus”, Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI 2012) Hong Kong and Shenzhen, China, 2-7 Jan 2012.

[2] Kamran Shaukat1,NayyerMasood,Sundas Mehreen1 and UlyaAzmeen ,” Dengue Fever Prediction: A Data Mining Problem”, Dar et al.J Data Mining Genomics Proteomics 2015.

[3] N.Subitha and Dr.A.Padmapriya “Diagnosis for Dengue Fever Using Spatial Data Mining”,International Journal of Computer Trends and Technology (IJCTT) ,August 2013.

[4] Daranee, Pratap Suriyaphol and Nuanwan,” Data Mining of Dengue Infection Using Decision Tree”,July 2015.

[5] cLambodar Jena, Narendra Ku. Kamila,” Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease”, International Journal of Emerging Research in Management &Technology, November 2015.

[6] Ahmed A.A.Radwan and TarekAbd El-Hafeez,” An Analysis of hepatitis C virus prediction using different data mining techniques”,December 2013.Reecha P. Yadav, Vinuchackravarthy Senthamilarasuand Krishnan Kutty, Sunita P. Ugale,".

[7] Huang, H. et al. “Business rule extraction from legacy code”, Proceedings of 20th International Conference on Computer Software and Applications, IEEE COMPSAC’96, 1996, pp.162-167

[8] Aarti Sharma,Rahul Sharma,Vivek Kr. Sharma,Vishal Shrivatava, “Application of Data Mining – A Survey Paper”, Department of CS & IT, A.C.E.I.T., Jaipur International Journal of Computer Science and Information Technologies- 2014.

[9] Arun George Eapen, “Application of Data mining in Medical Applications”, University of Waterloo, Waterloo, Ontario, Canada, 2004.

[10] Irshad Ullah, “Data Mining Algorithms And Medical Sciences”, International Journal of Computer Science & Information Technology (IJCSIT), Vol 2, No 6, December 2010