INVESTIGATION OF IMMUNIZATION PATTERN CHANGES WITH BIG DATA ANALYTICS IN ACCORDANCE WITH A RURAL AREA OF COIMBATORE REGION USING DECISION TREE ALGORITHM

M. Maheswari[1], T. Sumathi[2], M. Hemalatha[3]

ABSTRACT

Immunizations are made an important and integral part of the health development efforts. To improve the perceptions on immunization in a human, it is important to analyze its present status in those fields. This paper analyzes the perceptions on immunization in a rural area of Coimbatore. It is obtained by collecting and analyzing the rural data about immunization. This kind of data can be of large volume and that must be processed and transformed into useful results and hence data mining can greatly improve this task. The application of data mining techniques can produce important results in this analysis. The analysis of immunization awareness helps to know the percentage of people who are aware of immunization being provided. This can be used by government authorities to find areas that are needed to be improved in that rural area. The results obtained from this analysis can helps to know about the percentage of immunization awareness and improve it in rural areas. As data mining is the appropriate field to apply on this kind of big data, and knowledge extracted using data mining approaches will be useful to support government authorities. This project categorizes and analyzes the awareness among various immunization methods being adopted to the public welfare.

Keywords: Big data, immunization, classification, decision tree.

I. INTRODUCTION

[1] Data Mining is a detailed process of analyzing large amounts of data and picking out the relevant information. It refers to extracting or mining knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, accurate, manage, and process data within a tolerable elapsed time.http://en.wikipedia.org/wiki/ Big_data - cite_note-Editorial-13 Big data makes use of some techniques and technologies to acquire new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and are of a massive scale. Immunization is the process by which an individual's immune system becomes fortified against an agent (known as the immunogen). When this system is exposed to molecules that are foreign to the body, called *non-self*, it will orchestrate an immune response, and it will also develop the ability to quickly respond to a

[1]Department of Computer Science, Karpagam University, Coimbatore.
[2]Department of Computer Science, Karpagam University, Coimbatore.
[3]Department of Computer Science, Karpagam University, Coimbatore.

subsequent encounter because of immunological memory. This is a function of the adaptive immune system. Therefore, by exposing an animal to an immunogen in a controlled way, its body can learn to protect itself: this is called active immunization. The most important elements of the immune system that are improved by immunization are the T cells, B cells, and the antibodies B cells. Memory B cells and memory T cells are responsible for a swift response to a second encounter with a foreign molecule. Passive immunization is when these elements are introduced directly into the body, instead of when the body itself has to make these elements.

Immunization analysis is a task that finds the model of the perceptions on immunization status of people living in any rural area. This high volume of datasets and complexity of relationships between these kinds of data have made perceptions on immunization analysis an appropriate field for applying data mining techniques. Regional authorities everywhere have been handling a large amount of information and huge volume of records. Almost all regional authorities use the system to store and retrieve the awareness of the immunization details. This kind of big data has high volume and variety of data from which it is difficult to obtain useful results. Analyzing immunization data by applying data mining techniques provide us a technique for retrieving useful information that helps to know about a rural immunization awareness in a particular area. All the challenges occurred in handling this data has motivated this paper to focus on finding classification model. This paper classifies the awareness on immunization which can be used by the government authorities to improve the awareness based on immunization in that particular rural area.

Development of a region requires not only the data but also the useful information. Since data mining is the tool for extracting hidden patterns, the data mining techniques can be applied for analyzing the immunization dataset. Analyzing big data is an effective method in data mining to extract useful information. The objective of this project is to find the perceptions on immunization model in a rural. This project helps the regional authorities to improve facilities in the required field in a rural.

II. Related Works

Mrs. M.S.Mythili, Dr.A.R.Mohamed shanavas[1] analyzed the students performance by appling data mining classification algorithm in weka tool and applied various classification algorithm to find out the students performance. P.Yasodha, M.kannan[4] Analyze the records of diabetic patients to find out the characteristics that determine the presence of diabetes and to track the maximum number of men and women suffering from diabetes. Mrs.P.Nancy, Dr R.Keetharamani[7] compared the performance of data mining algorithm in classification of social network data and analyzed the impact of the internet on social group activities using data mining techniques. Smitha and suresh kumar[9] presented the applications of big data in data mining. Big data concepts can be used to implement data mining concepts. Milan Kumari, Sunila Godara[11] data mining classification techniques RIPPER classifier, Decision Tree, Artificial neural networks (ANNs), and Support Vector Machine (SVM) are analyzed on cardiovascular disease dataset. Performance of these techniques is compared through sensitivity, specificity, accuracy, error rate, True Positive Rate and False Positive Rate. S. M. Kamruzzaman[12] a

new algorithm for text classification using data mining that requires fewer documents for training. Instead of using words, word relation i.e. association rules from these words is used to derive feature set from pre-classified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of Genetic Algorithm has been added for final classification. Samir Kumar Sarangi, Dr. Vivek Jaglan, Yajnaseni Dash [13] focuses on an ongoing development and research activities of classification and clustering techniques for data mining and provides a review of machine learning algorithms used in data mining. A.Shameem Fathima, ,D.Manimegalai and Nisar Hundewale [14] summarizes various review and technical articles on arboviral diagnosis and prognosis. In this paper we present an overview of the current research being carried out using the data mining techniques to enhance the arboviral disease diagnosis and prognosis. This paper is not intended to provide a comprehensive overview of medical data mining but rather describes some areas which seem to be important from our point of view for applying machine learning in medical diagnosis for our real viral dataset.

III. Classification

[16] There are two forms of data analysis that can be used for extract models describing important classes or predict future data trends. These two forms are as follows:

- **Classification**
- **Prediction**

These data analysis help us to provide a better understanding of large data. Classification predicts categorical and prediction models predicts continuous valued functions. For example, we can build a classification model to categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation.

IV. Decision Tree Algorithm

[5]Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. In data mining, decision trees can be described also as the combination of mathematical and computational techniques to aid the description, categorization and generalization of a given set of data.

Data comes in records of the form:

$$(\mathbf{x}, Y) = (x_1, x_2, x_3, ..., x_k, Y)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The

vector x is composed of the input variables, x_1, x_2, x_3 etc., that are used for that task.

Decision tree learning is the construction of a decision tree from class-labeled training tuples. A decision tree is a flow-chart-like structure, where each internal (non-leaf) node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf (or terminal) node holds a class label. The topmost node in a tree is the root node.

PSEUDO CODE:

Algorithm : Generate_decision_tree

Input:

Data partition, D, which is a set of training tuples and their associated class labels.

attribute_list, the set of candidate attributes.

Attribute selection method, a procedure to determine the splitting criterion that best partitions that the data tuples into individual classes. This criterion includes a splitting_attribute and either a splitting point or splitting subset.

Output:

A Decision Tree

Method

```
create a node N;
if tuples in D are all of the same class, C then
  return N as leaf node labeled with class C;
if attribute_list is empty then
  return N as leaf node with labeled
  with majority class in D;|| majority voting
apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;
if splitting_attribute is discrete-valued and
  multiway splits allowed then // no restricted to binary
trees
attribute_list = splitting attribute; // remove splitting
attribute
for each outcome j of splitting criterion
   // partition the tuples and grow subtrees for each
partition
  let Dj be the set of data tuples in D satisfying outcome
j; // a partition
  if Dj is empty then
    attach a leaf labeled with the majority
    class in D to node N;
  else
    attach the node returned by Generate
    decision tree(Dj, attribute list) to node N;
  end for
return N;
```

Method:

Algorithms for constructing decision trees usually work top-down, by choosing a variable at each step that best splits the set of items.

Gini impurity

Used by the CART (classification and regression tree) algorithm, Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity

can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

To compute Gini impurity for a set of items, suppose $i \in \{1, 2, ..., m\}$, and let f_i be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m}(f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2$$

Information gain

Used by the ID3, C4.5 and C5.0 tree-generation algorithms. Information gain is based on the concept of entropy from information theory.

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i$$

indices for which the split test is false, respectively.

VI. Dataset Description

This data set is collected in real time from Madukkarai a rural village in Coimbatore district. It contains 89 attributes and 250 records.

Attribute name	Attribute description	Type
Street_nme	Name of the street	Numeric
Exercise	Practicing exercise are not	Numeric
e_type	Name of the exercise	Numeric
hand_wash	Hand wash. The value either 0 or 1	Numeric
sick_stay	The value is either 0 or 1. Stay at home when sick	Numeric
Cleaners	Name of the cleaners	Numeric
surf_clean	Cleaning time	Numeric
Meditation	Practicing meditation or not	Numeric
sleeping_hrs	Sleeping hours	Numeric
mos_ctrl	Name of the mosquito killer	Numeric
Water_cme	Quantity of water	Numeric
reg_food	Whether Take food regular time or not	Numeric
food_reason	If no specify the reason	Numeric
Snacks	Name of the snacks	Numeric
Veg	Include vegetables in daily diet or not	Numeric

nut_food	Whether eat nutritional foods daily or not	Numeric
kt_cln	Kitchen cleaning	Numeric
Wtr_purify	Name of the water purifier	Numeric
Veg_wash	Whether wash vegetables before cooking	Numeric
Iron	Iron foods	Numeric
Calcium	Calcium foods	Numeric
VitaminC	vitamin C foods	Numeric
vitaminB6	Vitamin B6 foods	Numeric
vitaminA	Vitamin A foods	Numeric
Protein	Protein foods	Numeric
Imm_aware	Awareness about immunization	Numeric
Medium	Name of the medium	Numeric
Awr_pgm	Awareness program	Numeric
Imm_help	Help to others	Numeric
Travel_vac	Travel vaccine	Numeric
Imm_fac	Immunization facility	Numeric
Hosp_visit	Visit hospitals when you are sick	Numeric

Prefer_hos	Prefer hospital either government or private	Numeric
Hos_reason	Reason	Numeric
Med_pres	Take medicine without doctors prescription	Numeric
Reg_health	Regular health check	Numeric
Check_time	Regular health check time	Numeric
Child_check	Regular health check on child	Numeric
Gov_fac	Government hospital facilities	Numeric
Need_fac	Need any extra facilities	Numeric
Nme_fac	Name of the facility	Numeric
Free_check	Free health checkups provided by the hospitals	Numeric
Diabetes	Diabetes	Numeric
BP	Blood Pressure	Numeric
Thyroid	Thyroid	Numeric
Heart_atk	Heart attack	Numeric
Treatment	Take treatment or not	Numeric

Ch_age	Age of the child	Numeric
BCG	Name of the vaccine at birth time	Numeric
BthOPV	Name of the vaccine at birth time	Numeric
BthHepB	Name of the vaccine at birth time	Numeric
6Wopv	Vaccine name at 6 weeks	Numeric
6wHepB	Vaccine name at 6 weeks	Numeric
6wHib	Vaccine name at 6 weeks	Numeric
10wOPV	Vaccine name at 10 weeks	Numeric
10wDPT	Vaccine name at 10 weeks	Numeric
10wHepB	Vaccine name at 10 weeks	Numeric
10wHib	Vaccine name at 10 weeks	Numeric
14wOPV	Vaccine name at 14 weeks	Numeric
14wDPT	Vaccine name at 14 weeks	Numeric
14wHepB	Vaccine name at 14 weeks	Numeric
14wHib	Vaccine name at 14 weeks	Numeric
6mHepB	Vaccine name at 6 months	Numeric
9m_mss	Vaccine name at 9 months	Numeric
18mOPV	Vaccine name at 18 months	Numeric
18mDPT	Vaccine name at 18 months	Numeric
18mHib	Vaccine name at 18 months	Numeric
18mMMR	Vaccine name at 18 months	Numeric
2yTyphoid	Vaccine name at 2 years	Numeric
5yOPV	Vaccine name at 5 years	Numeric
5yDPT	Vaccine name at 5 years	Numeric
5yMMR	Vaccine name at 5 years	Numeric
10yTdap	Vaccine name at 10 years	Numeric
Vcc_reason	Reason	Numeric
Pblms	Problems when not vaccinated	Numeric
Int_vcc	Interested to get vaccine	Numeric
Prevent_vcc	Get vaccines to prevent from diseases	
W_age	Age of women	Numeric

W_age	Age of women	Numeric
Tw_BP	BP test for women	Numeric
Bone	bone test for women	Numeric
Breast	Breast test for women	Numeric
Tw_chol	Cholesterol test for women	Numeric
Tw_dia	diabetes test for women	Numeric
M_age	Age of the men	Numeric
Tm_chol	Cholerterol test for men	Numeric
Tm_BP	BP test for men	Numeric
Tm_dia	diabetes test for men	Numeric

7. RESULTS AND DISCUSSION

The above dataset has been classified based on various awareness criteria like exercise, hospitals, awareness, facilities, etc. The classification results obtained for exercise criteria is given below.

Exercise:

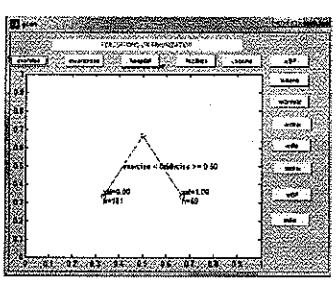

Fig 6.1 exercise criteria

Figure 6.1 shows only the 27.6% people have the habit of doing exercise regularly remaining 72.4% people don't have this habit in this rural area.

Immunization awareness:

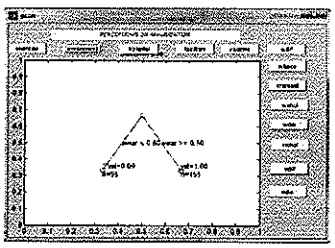

Fig 6.2 awareness of immunization

Figure 6.2 shows 62% people having the awareness about the immunization and 38% people not having the immunization awareness.

Hospitals

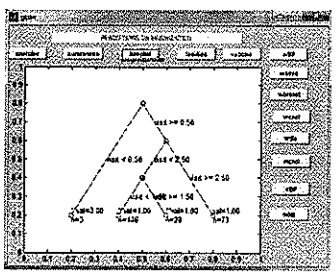

Fig 6.3 preferred hospitals

Figure 6.3 shows 1.2% people not interested to visit the hospitals when sick and 98.8% people interest to visit the hospitals. In this, 58.4% people prefer the private hospitals, 11.2% people visit the government hospitals and 29.2% people prefer the both government and private hospitals.

Facilities:

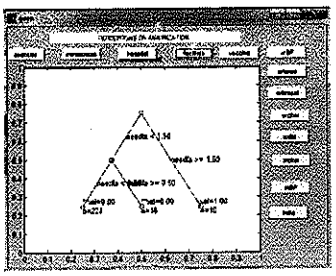

Fig 6.4 facilities criteria

Figure 6.4 shows the 6.4% people specify the facility is more staffs needed and 4% people needs scan facility is needed.

BP test for women

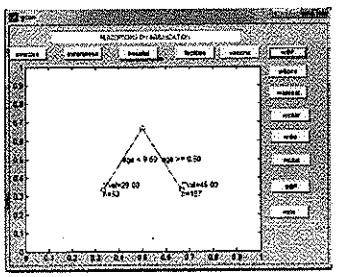

Fig 6.5 BP test for women

Figure 6.5 shows the 78.8% people have taken BP test and 21% people have not taken BP test.

Cholesterol test for women:

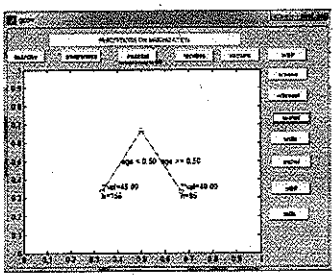

Fig 6.6 cholesterol test for women

Figure 6.6 shows 62% people have taken the cholesterol test and remaining 38% people have not taken the cholesterol test.

Breast test for women:

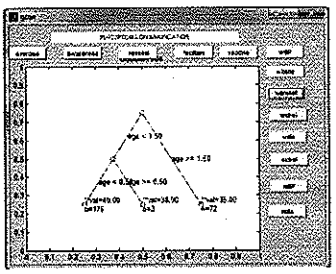

Fig 6.7 breast test for women

Figure 6.7 shows only 1.2% people have taken the breast cancer test, 28.8% people are not applicable for this test and 71% people do not taken this test in this rural.

Bone mineral density test for women:

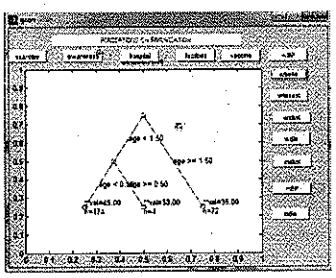

Fig 6.8 bone test for women

Figure 6.8 shows 1.6% people have taken the bone mineral density test,71% people are not applicable for this test and 68.4% people are do not taken this test.

Diabetes test for women:

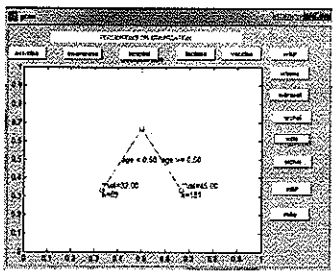

Fig 6.9 diabetes test for women

Figure 6.9 shows 72.4% people have taken the diabetes test and remaining 27.6% people have not taken the test.

Cholesterol test for men:

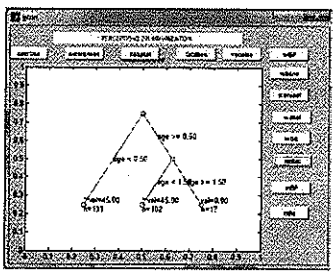

Fig 6.10 cholesterol test for men

Figure 6.10 shows 52% people are not taken the cholesterol test, 41% people have taken the cholesterol test.

BP test for men

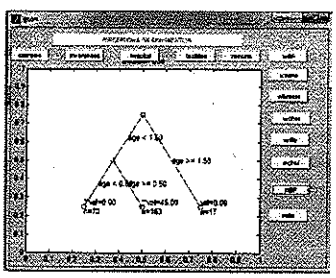

Fig 6.11 BP test for men

Figure 6.11 shows the 28% people are not taken the BP test, 65.2% people are taken the BP test.

Diabetes test for men:

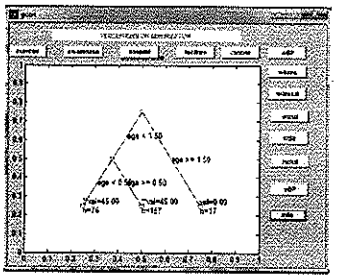

Fig 6.12 diabetes test for men

Figure 6.12 shows 62.8% people are taken the diabetes test, 31% people are not taken the this test.

CONCLUSION AND FUTURE ENHANCEMENT

The main aim of the research is to find out the truth which is hidden and which has not been discovered yet. In this paper, the answer to the questions has been discovered through the application of scientific procedures. The data set for the study contains Immunization data and it contains 89 Attributes 250 instances of this specific region. Decision tree classification algorithm is used to discover and understand the underlying patterns involved in the immunization data set. As a conclusion of this work the result obtained shows that only 62% people having the immunization awareness, 72.4% women have taken the BP test, 1.2% women have taken the breast cancer test, 1.6% women have taken the bone mineral density test, 62% women have taken the cholesterol test,72.4% women have taken the diabetes test, 52.4% men have taken the cholesterol test, 62.8% men have taken diabetes test and 65.2% men have taken the BP

test. In future this project can be applied in all other regions data. It can be applied to real life application such as in hospitals to collect data on what type of diseases the mostly people are suffering i.e. region wise which disease is mostly found in people. It can also include comparing with other algorithm of classification which can also be applied in real life applications.

References

[1] Harvinder Chauhan, Anu Chauhan, *"Implementation of decision tree algorithm c4.5"* International Journal of Scientific and Research Publications, Volume 3, Issue 10, October 2013 1ISSN2250-3153.

[2] Chen, H., Zhan, Y., Li, Y., (2010), *"The application of Decision Tree in Chinese email Classification"*. In the proceeding of 9th International Conference on machine Learning and Cybernetics, 2010, pp. 305-308.

[3] Jaiwei Han and Micheline Kamber, *"Data Mining Concepts and Techniques"*,

[4] Yashoda, P., Kannan, M., (2011), *"Analysis of a Population of Diabetic Patients Databases in Weka tool"*. In the International Journal of Scientific and Engineering Research Volume 2, 2011, pp. 1-5.

[5] Decision tree learning – Wikipedia, the free encyclopedia.

[6] Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas, *"An Analysis of students' performance using classification algorithms"*, IOSR Journal of Computer Engineering (IOSR-JCE),Volume 16, Issue 1, Ver. III (Jan. 2014), PP 63-69.

[7] Mrs.P.Nancy, Dr.R.Geetha Ramani, *"A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data"*,International Journal of Computer Applications (0975 – 8887) Volume 32– No.8, October 2011.

[8] P. C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis, Understanding big data – Analytics for enterprise class Hadoop and streaming data, McGraw-Hill, 2012.

[9] International Journal of Emerging Technology and Advanced Engineering ,(ISSN 2250-2459, ISO 9001:2008 Certified Journal) Application of Big Data in Data Mining Volume 3, Issue 7, July 2013.

[10] Ritika, *"Research on Data Mining Classification"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014.

[11] Milan Kumari, Sunila Godara, *"Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction"*, IJCST, Vol. 2, Issue 2, June 2011.

[12] S. M. Kamruzzaman, *"Text Classification Using Data Mining"*, ICTM ,2005.

[13] Samir Kumar Sarangi, Dr. Vivek Jaglan, Yajnaseni Dash, *"A Review of Clustering and Classification*

Techniques in Data Mining",International Journal of Engineering, Business and Enterprise Applications (IJEBEA).

[14] A.Shameem Fathima ,D.Manimegalai and Nisar Hundewale,*"A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue"*, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011 .

[15] Dorina Kabakchieva,*"Predicting Student Performance by Using Data Mining Methods for Classification"*, cybernetics and information technologies, Volume 13, No 1,2013.

[16] www.tutorialspoint.com

AUTHOR'S BIOGRAPHY

Dr. Sumadhi.T is an Associate Professor, Department of Computer Science in Karpagam Univesity, Coimbatore. She received her B.Sc.(CS), in 1996 and MCA in 1999 from Bharathiar University, Coimbatore. She obtained her M.Phil. in the area of Image Processing from Periyar University, Salem in 2006. Her research interest lies in the area of Image Processing and Data mining. . She obtained her Ph.D. in the area of Image Processing from Karpagam University, Coimbatore in 2013. She has published about 20 papers in international and national journals. She has attended about 15 international and national conference held at various places in Tamil Nadu.

Dr. M.Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University,. Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science in Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.