www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume - 3 Issue -9 September, 2014 Page No. 8236-8241

Identification Of Malnutrition With Use Of Supervised Datamining Techniques –Decision Trees And Artificial Neural Networks

D.Thangamani [*] ***P.Sudha*** [#]

[*]*Research Scholar of Computer Science,* [#] *Assistant professor, Department of Computer Science,*
Sree Saraswathi Thyagaraja College
Pollachi – 642 107, Coimbatore, Tamil Nadu, India
[*]Email: thangamvelu@gmail.com
[#]Email: sudha_sabariananth@yahoo.co.in

Abstract **--**In today's modern world, Globalization, demographic transition, life style changes and dietary meal patterns influences people's nutrition. This work attempts to demonstrate the analysis of malnutrition based on food intakes, wealthy index, age group, education level, occupation, etc. Objective of this work is to use of effective supervised machine learning techniques-decision trees and artificial neural networks to classify dataset of family health survey and Classification and prediction techniques provides appropriate and flexible methods to process large amount of data for specifying accurate malnutrition detection and prevention over the survey dataset. The result of supervised data mining techniques in nutrition database provides the nutrition status of children age under five. This work is useful to improve nutrition level of public health with the help of government health services to the people.

Keywords : malnutrition, decision trees, artificial neural networks

INTRODUCTION

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large database. The patterns must be actionable so that they may be used in enterprise's decision making process[4].

Data mining or knowledge discover in databases (KDD)is a collection of exploration techniques based on advanced analytical methods and tools for handing for large amount of information. The techniques can find novel patterns that may assist an enterprise in understanding the business better and in forecasting. Many data mining techniques are closely related to some of the machine learning techniques that have been developed over the last 40 years. Others are related to techniques that have been developed in statistics, sometimes called exploratory data analysis. These techniques were developed some time ago and were designed to deal with a limit amount of data. The techniques have now been modified to deal with large amounts of data[2].

This attempt using supervised machine learning techniques-decision tree and artificial neural network. Decision tree is a tree based knowledge representation methodology used to represent classification rules. The leave nodes represent class labels while other nodes represent the attributes associated with the objects being classified the branches of tree represent each possible value of the attribute value from which it originates. Decision tree are grown through an iterative splitting of data into discrete groups. Decision tree which are used to predicate categorical variables are called classification tre\ because it plays instances in categories. Decision tree used to predicate continuous variable are called regression tree. Some of the decision tree are ID3, C5.0, Quest, CART and CHAID. The main advantages of decision tree are execution efficiency mainly due to its simple and economical representation and its ability to perform.

Neural network is defined as a data processing system consisting of large number of simple highly interconnected processing elements in an architecture inspired by the structure of brain. There are three types of ANNs are Single layer feed forward network, Multi layer feed forward Network and recurrent network.

This attempt analyzes the status of children's nutrition based on their intakes of foods. Malnutrition prediction can be made by using following children's diet chart. A balanced diet is essential for children of all ages. Parents must ensure that children eat meals consisting of all food groups to ensure healthy children. A balanced diet consists of carbohydrates, proteins, vitamins and minerals and also meets the daily caloric needs of the body. This

means, 50% of your calorie needs should be derived from carbohydrates, 20% from protein and 30% from fats. Children should be provided daily, with a diet consisting of all the above mentioned vitamins and minerals. Therefore, some balanced diet charts for children are listed here.

Nutrition	Food Group	Recommended %
Carbohydrates	Cereals and grains, etc.	33%
Vitaminsand Minerals	Various fruits and vegetables	33%
Meat Protein	Fish, meat and eggs	12%
Milk Proteins	Dairy products	15%
Fat and Sugar	Fatty foods, sugary sweets etc.	7%

Table 1 Children Diet Chart

RELATED WORK

In [6],Myonghwa Park, Hyeyoung Kim,Sun Kyung Kim, proposed C5.0, C&R Tree, QUEST, and CHAID models, the highest predictability was reported by C&R Tree with the accuracy rate of 77.1%. The presence of more than two co morbidities, living alone status, having severe difficulty in daily activities, and lower perceived economic status were identified as risk factors of malnutrition in elderly. They gave a reliable decision support model was designed to provide accurate information regarding the characteristics of elderly individuals with malnutrition. The findings demonstrated the good feasibility of data mining when used for a large community data set and its value in assisting health professionals and local decision makers to come up with effective strategies for achieving public health goals.

In [7], Xu Dezhi, Gamage Upeksha Ganegoda et al, proposed rule based classification is used alone with Agent Technology to detect malnutrition in children**.** Detection Agent will be enhanced by introducing rule based classification techniques. To implement the experiment the database consists of 200 records. From that 150 records will be used as training set and 50 records as test set. According to the questions given to the users a set of rules will be generated. Then for each rule, (1) likelihood ratio value will be calculated. Then the system will select the best 35 rules according to the likelihood value. By using the 35 rules it will decide the final answers of the test set. Therefore it can compare the actual value with the obtain value.

In [8], Aine P Hearty and Michael J Gibney, proposed data mining techniques to the food and meal-based coding systems. The ANN had a slightly higher accuracy than did the decision tree in relation to its ability to predict HEI quintiles 1 and 5 based on the food coding system (78.7% compared with 76.9% and 71.9% compared with 70.1%, respectively).However, the decision tree had higher accuracies than did the ANN on the basis of the meal coding system (67.5% compared with 54.6% and 75.1% compared with 72.4%, respectively).

METHODS

A Survey conducted by Indian Institute of Population Science, Mumbai, India named NFHS-III(National Family Health Survey) to collect all the health related data. In Tamil Nadu data were collected by Gandhi gram rural institute, Dindigul. In this work 254 child data of age below five can used to show nutritional status of children. This classification and rules can follows nutritional recommendation. These recommendation considers anemia level, Breast feed, vegetables, fruits, nuts, grain foods, milk products, tinned powder, baby cereal, etc. Dataset consists of 51555 data objects including all states, in this Tamil Nadu dataset 1736 data object. This work utilized 254 child data objects. This application implemented in weka-3.6.11 version.

DATAMINING TECHNIQUES

DECISION TREE MODELS

Decision tree learning is a method commonly used in data mining.[1] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node responds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

There are many specific decision-tree algorithms. Such algorithms are (i)ID3 (Iterative Dichotomiser 3), (ii)C4.5 (successor of ID3), (iii)CART (Classification And Regression Tree) ,(iv)CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees[1](v)MARS: extends decision trees to better handle numerical data.(vi)Conditional Inference Trees. Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid over fitting. This approach results in unbiased predictor selection and does not require pruning.

ID3 ALGORITHM

ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains.

Algorithm

The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy $H(S)$ (or information gain $IG(A)$) of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set S is then split by the selected attribute (e.g. age < 50, 50 <= age < 100, age >= 100) to produce subsets of the data. The algorithm continues to recurs on each subset, considering only attributes never selected before.

RANDOM FOREST TREE

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Leo Breiman and Adele Cutler, and "Random Forests" is their trademark.

Algorithm

The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1$ through y_n, bagging repeatedly selects a bootstrap sample of the training set and fits trees to these samples:

For $b = 1$ through B:

- Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
- Train a decision or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x')$$

or by taking the majority vote in the case of decision trees.

In the above algorithm, B is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. Increasing the number of trees tends to decrease the variance of the model, without increasing the bias. As a result, the training and test error tend to level off after some number of trees have been fit. An optimal number of trees B can be found using cross-validation, or by observing the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample.

ARTIFICIAL NEURAL NETWORKS

An artificial neural network (ANN) is a computational model that attempts to account for the parallel nature of the human brain. An (ANN) is a network of highly interconnecting processing elements (neurons) operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. A subgroup of processing element is called a layer in the network. The first layer is the input layer and the last layer is the output layer. Between the input and output layer, there may be additional layer(s) of units, called hidden layer(s). If a neural network needs to train to perform a particular function by adjusting the values of the connections (weights) between elements.

All artificial neural networks are divided into two learning categories: supervised and unsupervised. In supervised learning, the network is trained by providing it with input and output patterns. During this phase, the neural network is able to adjust the connection weights to match its output with the actual output in an iterative process until a desirable result is reached. An ANN of the unsupervised learning type, such as the self-organizing map, the neural network is provided only with inputs, there are no known answers. The network must develop its

A multilayer perceptron (MLP) is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. A MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable

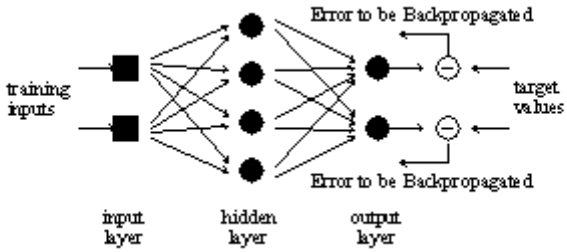

Figure 1 Multilayer backprobagation

Learning through backpropagation

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This is an example of supervised learning, and is carried out through backpropagation, a generalization of the least mean squares algorithm in the linear perceptron.

It represent the error in output node j in the nth data point by $e_j(n) = d_j(n) - y_j(n)$, where d is the target value and y is the value produced by the perceptron. We then make corrections to the weights of the nodes based on those corrections which minimize the error in the entire output, given by

$$\mathcal{E}(n) = \frac{1}{2}\sum_{j} e_j^2(n)$$

.

Using gradient descent, we find our change in each weight to be

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

where y_i is the output of the previous neuron and η is the *learning rate*, which is carefully selected to ensure that the weights converge to a response fast enough, without producing oscillations. In programming applications, this parameter typically ranges from 0.2 to 0.8

The derivative to be calculated depends on the induced local field v_j, which itself varies. It is easy to

prove that for an output node this derivative can be simplified to

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = e_j(n)\phi'(v_j(n))$$

where ϕ' is the derivative of the activation function described above, which itself does not vary. The analysis is more difficult for the change in weights to a hidden node, but it can be shown that the relevant derivative i

$$-\frac{\partial \mathcal{E}(n)}{\partial v_j(n)} = \phi'(v_j(n))\sum_k -\frac{\partial \mathcal{E}(n)}{\partial v_k(n)} w_{kj}(n)$$

.

This depends on the change in weights of the kth nodes, which represent the output layer. So to change the hidden layer weights, we must first change the output layer weights according to the derivative of the activation function, and so this algorithm represents a backpropagation of the activation function.

RESULTS

Classification is the most familiar data mining technique estimation and prediction may be viewed as types of classification. Classification maps data into predefined attributes shown in figure. The variables are anemia level, Breast feed, Baby Cereals, Grain foods, Fruits/ Juice, Green vegetables, fish/ meat, non-veg / egg, tinned powder/fresh groups or classes. It is referred as supervised learning because the classes are determined before examines the data. Prediction can be classifying attributes into one of a set of possible classes.

Input variables and Attribute selection:

Classification techniques used totally 28 attributes of child id, state { Tamil nadu }, Age {0,1,2,3,4},Sex { Male , Female }, Religion { Hindu , Christian , Muslim } , Mothers Education' { No Education , Secondary , Primary , Higher } Wealth Index { Middle , Poor , Rich} Anemia Level{ Not anemic , Mild , Moderate , Severe }, Breast Feed{ Yes , no }, Plain Water{ Yes , no }, Juice { no , Yes },Tea/ Coffee{ Yes , no }, Tinned powder/ Fresh Milk{ Yes , no },Baby Cereal { no , Yes }, non-Veg{ no , Yes }, peens /peas food { no , Yes }, Nuts{ no , Yes } Grain foods { Yes , no } ,Potato/cassava { no , Yes }, Egg{ no , Yes }, Green Vegetables { no , Yes }, Fruits{ no , Yes } , Fish Foods' { no , Yes }, Food from Nuts { no , Yes } Milk Products' { no , Yes } , Oil/Fat Foods { no , Yes }, Solid/Semisolid Foods { no , Yes } , Iodized Salt { No Iodine , No Quality }

Selecting Important Variables :

As a result of attribute selection, 9 important variables were identified among total 28

milk were identified and applied to the rules specified by food recommendation to show status of child nutrition level.

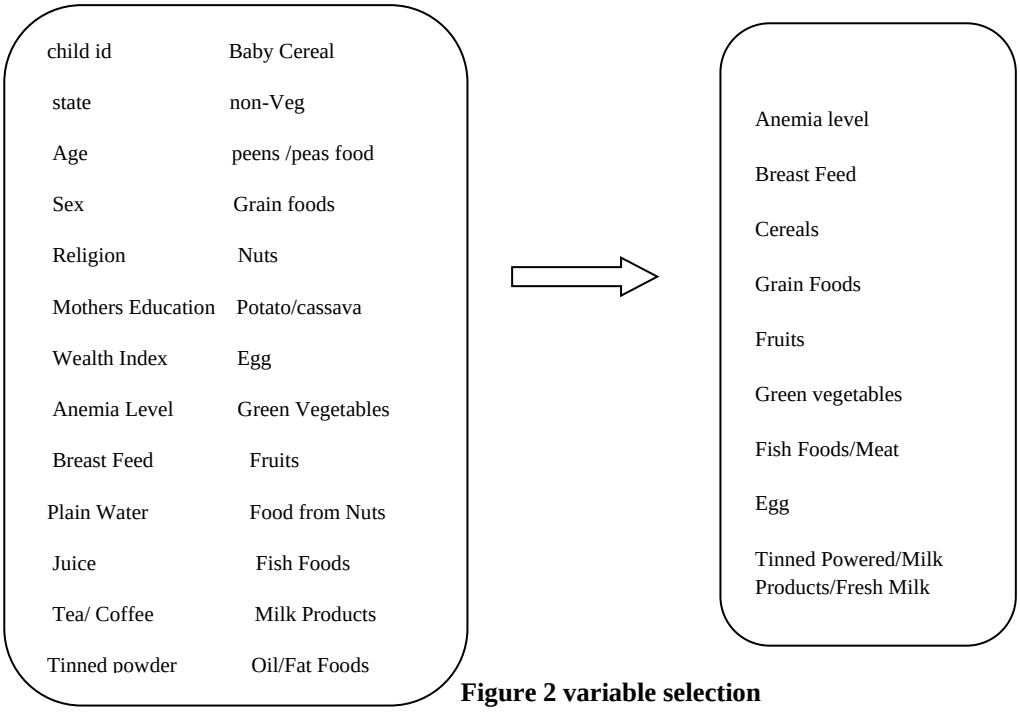

Figure 2 variable selection

Result of general characteristic:

A total of 254 child included in this work. This sample consisted five category of age. Age 0(below 1 year) had 91.53% poor nutritional state, while 8.47% people had normal nutritional status. Age of 1 year child had 57.53% poor nutritional status. Age of 2 years child had poor nutritional status 66.22%. Age of 3 years child had poor nutritional status 61.90%. Age of 4 years had poor nutritional status 81.48%. Gender of Male had 90.90% of malnutrition. Gender of Female had 97% of malnutrition. Religion of Hindu had 75.22% of poor nutritional status. Religion of Christian had 59.09% of poor nutritional status. Religion of Muslim had 30% of poor nutritional status. Mothers Education of Non-Education had 82.98% of poor nutritional status. Mothers Education of Secondary had 69.94% of poor nutritional status. Mothers Education of Primary had 69.04% of poor nutritional status. Mothers

Education of Higher had 52.39% of poor nutritional status. Wealth Index of poor had 86.27% of poor nutritional status. Wealth Index of Middle had 75.31% of poor nutritional status. Wealth Index of rich had 61.47% of poor nutritional status.

This result displayed in Table 2.

Result of important variables:

Anemia level of Not Anemic had 57.66% of malnutrition and Moderate/severe had 78.32% of malnutrition. The Child who had breast feed had 69.93% of malnutrition and the child who had not breast feed had 70.25% of malnutrition. The child who had baby cereals had 62.90% of malnutrition. The child who had not had baby cereals had 72.40% of malnutrition. The child who had grain foods daily/weekly had 61.45% of malnutrition. The child who had grain foods occasionally or never had in 80% of malnutrition. The child of 50% malnutrition who had fruit juice daily or weekly. The child of 79.21% of malnutrition who had fruit juice occasionally or never. The child of 45.71% of malnutrition who had green vegetables daily or weekly. The child of 73.97% of malnutrition who had green vegetables occasionally or never. The child of 43.75% of malnutrition who had fishes and meat daily or weekly. The child 76.21% of malnutrition who had fishes and meat occasionally or never. The child of 63.64% of malnutrition who had non-veg daily or weekly. The child of who had non-veg, occasionally or never. Table 3 shows above results.

Results to show nutritional level

	Normal	Malnutrition	Total
Anemia Level			
Not Anemic	47 (42.34%)	64 (57.66%)	111
Moderate/severe	31 (21.68%)	112 (78.32%)	143
Breast Feed			
Yes	40 (30.08%)	93 (69.93%)	133
No	36 (29.75%)	85 (70.25%)	121
Baby Cereals			
Yes	23 (37.10%)	39 (62.90%)	62
No	53 (27.60%)	139 (72.40%)	192
Grain Foods			
Daily/weekly	69 (38.55%)	110 (61.45%)	179
Occasionally/never	15 (20%)	60 (80%)	75
Fruits/Juice			
Daily/weekly	38 (50%)	38 (50%)	76
Occasionally/never	37 (20.79%)	141(79.21%)	178
Green Vegetables			
Daily/weekly	19 (54.29%)	16 (45.71%)	35
Occasionally/never	57 (26.03%)	162 (73.97%)	219
Fishes/Meat			
Daily/weekly	27 (56.25%)	21 (43.75%)	48
Occasionally/never	49 (23.79%)	157 (76.21%)	206
Non-veg/Egg			
Daily/weekly	12 (36.36%)	21 (63.64%)	33
Occasionally/never	62 (28.05%)	159 (71.95%)	221
Tinned Powder/ Fresh Milk			
Daily/weekly	61 (38.36%)	98 (61.64%)	159
Occasionally/never	14 (14.74%)	81 (85.26%)	95

Table 3 variables used final modeling

Predicting performance according to classification techniques:

Datasets were applied to different classification techniques which provide different accuracy, time to build model, error rate. The following tables and charts provide performance results of different classification techniques.

	Normal	Malnutrition	Total
Age			
0	5(8.47%)	54(91.53%)	59
1	31(42.47%)	42(57.53%)	73
2	25(33.78%)	49(66.22%)	74
3	8(38.10%)	13(61.90%)	21
4	5(18.52%)	22(81.48%)	27
Gender			
Male	39(29.10%)	95(70.90%)	134
Female	36(3%)	84(97%)	120
Religion			
Hindu	55(24.78%)	167(75.22%)	222
	9(40.91%)	13(59.09%)	22
Christianity			
Muslims	7(70%)	3(30%)	10
Mothers Education			
Non-Education	8(17.02%)	39(82.98%)	47
	44(30.56%)	100(69.44%)	144
Secondary			
Primary	13(30.96%)	29(69.04%)	42
Higher	10(47.61%)	11(52.39%)	21
Wealth Index			
Poor	7(13.73%)	44(86.27%)	51
Middle	20(24.69%)	61(75.31%)	81
Rich	47(38.53%)	75(61.47%)	122

Table 2 general charact eristics

The highest percentage of accuracy shown by multilayer perceptron network in table. The result of all three id3, random forest, Multilayer perceptron neural network accuracy, TP rate, FP rate, Error rate, Time to take build model are differ by each attributes(class) of data. Table 4 shows result of the class anemia level, class iodized salt.

Model	Accuracy	CVE rate	Time(seconds)
ID3	68.50 %	31.50%	0.03
Random Forest	77.17%	22.83%	0.02
Multilayer Perceptron	77.17%	22.83%	5. 77

Table 4 shows highest accuracy percentage of multilayer perceptron 77.17%, lowest error rate 22.83% and time 5.77 seconds. It shows same accuracy with random forest tree(77.17%), lowest error rate 22.83%. and time 0.02 seconds.

Table 4 performance comparison

CONCLUSION

Prediction can be classifying attributes into one of a set of possible classes. The use of decision trees and neural networks to classify dataset. Double Burdon of malnutrition analyzed and identified by use of database with supervised data mining techniques. Many children have poor eating habits, which can lead to various long-term health complications, such as obesity, heart disease, type 1 diabetes and osteoporosis. Ensuring that people's child learns the importance of eating a balanced diet, means ensuring he or she is free of these diseases and grows up to be a healthy adult. This study will help to the people and the health care professionals to improve the nutrition of the people. This study will also work to identify those people which needed special attention to reduce deficiency and overdose and taking appropriate action for the nutrition

improvements. Future work will identify various nutrition style, people's information from the use of NFHS(2014 - 2015) –IV database.

ACKNOWLEDGEMENTS

The author would like to thank DHS(Demographic and Health Surveys), IIPS(Indian Institute of Population Science) for give authorization to download NFHS III(National Family Health Survey) dataset and Assistant Professor Mrs.P.Sudha for her guidance in carrying out this work. Thanks are also to Gandhi gram University, Dindigul authorities for providing guidance to know about NFHS. On the personal front, the author is grateful to her family for their support and motivation in doing research work.

REFERENCES

[1] Margaret H. Dunham and S. Sridhar, "Data Mining Introductory and Advanced Topics", ISBN: 0130888923, Pearson Education, Inc., Copyright 2003.
[2] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques: Concepts and Techniques", third edition.
[3] S.Rajasekaran and G.A .Vijayalakshmi Pai, Neural Networks ,Fuzzy logic and genetic algorithms Synthesis and Applications,PHI Learning Private Limited,2003.
[4] K.P.Somen, Shyam Diwakar and V,Ajay, "Insight into Data Mining Theory and practice", Prentice hall of India Private Limited,2008.
[5] Quinlan, C4.5 Programs for Machine Learning, Morgan Kaufmann publishers, 1993. [6] Myonghwa Park, PhD, Hyeyoung Kim, PhD, Sun Kyung Kim, MSN, " Knowledge Discovery in a Community Data Set: Malnutrition among the Elderly "**,** Healthcare Informatics and Research,2014.
[7] Xu Dezhi and Gamage Upeksha Ganegoda , "Rule Based Classification to Detect Malnut,rition in Children**",** International Journal on Computer Science and Engineering (IJCSE),2011.
[8] Aine P Hearty and Michael J Gibney, **"**Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees", American Society for Nutrition", 2008.
[9] Qeethara Kadhim Al-Shayea ,**"** Artificial Neural Networks in Medical Diagnosis", IJCSI International Journal of Computer Science,2011.
[10] Jameela Ali Akrimi, Abdul RahimAhmad, Loay E. George(IJSR), "Review of Machine Learning Techniques in Anemia Recognition", International Journal of Science and Research (IJSR), India Online ISSN: 2319-7064, 2013.
[11] K.R. Lakshmi , M.Veera Krishna and S.Prem Kumar," Performance Comparisonof Data Mining Techniques for Predicting of Heart Disease Survivability"**,** International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 ISSN 2250-3153.
[12] M. Goebel and L.Gruenwald, A survey of data mining and knowledge discovery software tools, SIKDD Explorations, vol. 1, Issue 1, pp.22-33, June 1999.
[13] Carling, A. (1992). Introducing Neural Networks. Wilmslow, UK: Sigma Press.
[14] Fausett, L. (1994). Fundamentals of Neural Networks. New York: Prentice Hall.
[15] Haykin, S. (1994). Neural Networks: A Comprehensive Foundation. New York: Macmillan Publishing. [16] Patterson, D. (1996). Artificial Neural Networks. Singapore: Prentice Hall. [17] www.dhsprogram.com [18] www.iipsindia.org [19] www.MDhealth.com

AUTHORS

D.Thangamani has completed M.Sc (CS & IT), She is currently doing research in Data Mining and Knowledge Discovery (M.Phil) in Sree Saraswathi Thyagaraja College, Pollachi. Tamilnadu, India.
Email: thangamvelu@gmail.com

P.Sudha has received MCA., Mphil.,presently she has working as a Assistant professor, in PG Department of Computer Science, Sree Saraswathi Thyagaraja College at Pollachi – 642 107, Coimbatore, Tamil Nadu, India.
Email: sudha_sabariananth@yahoo.co.in.