International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

PREDICTION OF HEART DISEASE USING RNN ALGORITHM

N. Sowri Raja Pillai [1], K.Kamurunnissa Bee[2], J.Kiruthika[3]

[1] *N. Sowri Raja Pillai, Professor, HOD of Information Technology, ACETW, Pondicherry, India*
[2] *K. Kamurunnissa Bee, Dept. of Computer Science and Engineering, ACETW, Pondicherry ,India*
[3] *J. Kiruthika, Dept. of Computer Science and Engineering, ACETW, Pondicherry ,India*

Abstract *- An infrastructure builds in the data mining platform which is reliable to challenge the commercial and non- commercial IT development communities of data streams in high dimensional data cluster modeling. The knowledge discovery in database (KDD) is alarmed with development of methods and techniques for making use of data. The data size is generally growing from day to day. One of the most important steps of the KDD is the data mining which is ability to extract useful knowledge hidden in this large amount of data. Both the data mining and healthcare industry have emerged some of reliable early Heart diseases detection systems and other various healthcare related systems from the clinical and diagnosis data. In this project we propose the enhanced data mining algorithm for healthcare application. This proposed method is a critical issue to predict the heart disease diagnosis of adult disease patients due to the possibility of spreading to high-risk symptoms in medical fields. Most studies for predicting prognosis have used complex data from patients such as biomedical images, biomarkers, and pathological measurements. We demonstrate a language model-like method for predicting high-risk prognosis from diagnosis histories of patients using deep recurrent neural networks (RNNs), i.e., prognosis prediction using RNN (PP-RNN). The proposed PP-RNN uses multiple RNNs for learning from diagnosis code sequences of patients in order to predict occurrences of high-risk diseases. Finally our experimental result shows our proposed method can achieve more accuracy result.*

Keywords: Heart disease; Data mining; Data Cleaning, Clustering, RNN- algorithm.

1. INTRODUCTION

Healthcare systems are highly complex, fragmented and use multiple information technology systems. With vendors incorporating different standards for similar or same systems, it is little wonder that all-round inefficiency, waste and errors in healthcare information and delivery management are all too commonplace an occurrence. Consequently, a patient's health records often get trapped in silos of legacy systems, unable to be shared with members of the healthcare community. These are some of the several motivations driving an effort to encourage standardization, integration and electronic information exchange amongst the various healthcare providers. The study termed as Developmental Origins of Health and Diseases or DOHAD has successfully proven the importance of developmental records of individuals in predicting and or explaining the diseases that a person is suffering from. In the current largely paper-based health records world, invaluable data is more often than not unavailable at the right time in the hands of the clinical care providers to permit better care. This is largely due to the inefficiencies inherent in the paper-based system. In an electronic world, it is very much possible, provided certain important steps are taken beforehand, to ensure the availability of the right information at the right time. Supervised learning is the machine learning task of interfering a function from labeled training data. We are using supervised learning for training set in this project.

Ensemble of classifiers is combinations of multiple classifiers, referred as base classifiers. Ensembles usually achieve better performance than any of the single classifiers. In order to build a good base classifiers, also the base classifiers must be diverse, this means that for the same instance, the base classifiers return different outputs and their errors should be in different instances.

1.1 RELATED WORKS

"Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M"[1]. Heart Disease Diagnosis Using Data Mining TechniqueIn this paper, using only 14 attributes example age , gender,bmi,sugar , down sloping and fat. Less accuracy

"Jyoti Soni"[2] Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction, this paper using Naive Bayes, Decision List and KNN. Here the Classification based on clustering is not performing well.

S.N.Deepa and B.Aruna Devi, "Artificial Neural Networks design for Classification of Brain Tumour".In this system, we manipulate the ability of Back Propagation Neural Networks (BPN) and Radial Basis Function Neural network (RBFN) to categorize mind MRI pictures to either cancer or noncancerous growth instantly. The results revealed outperformance of RBFN criteria in comparison to BPN with category precision of 85.71% which performs as appealing device for category and needs expansion in mind growth analysis.

2. EXISTING SYSTEM

Our existing system uses breast cancer datasets which has two classes' recurrence events and no recurrence events. Pre-process the data and then classifying the data using decision stump which is to be used as a base classifier for AdaBoost algorithm with number of iteration is set to 2 and weight threshold for weight pruning is set to 10 AdaBoost. M1 algorithm is used, which use the base classifier Decision Stump(AdaBoost_DS) and reweighting, the number of iterations is set on 10, and weight threshold for weight pruning is set on 100.Comparing the correctly classified instance and accuracy of classification ,AdaBoost implementation of decision stump improves accuracy. AdaBoost, short for "Adaptive Boosting", is a machine learning Meta algorithm; It can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. On robust error a norm, the ada booster technique rule has diffusion is equivalent to regularization with an explicit line process. Removing the noise is a critical technology of wavelet image processing: the wavelet transform can decompose the data processing into characteristic coefficients with different resolutions and the latter can then be analyzed and processed in order to remove noise.

2.1 Genetic algorithm

A genetic algorithm (GA) is a searching that imitate the process of natural evolution. This inquisitive is routinely used to generate useful solutions to optimization and search problems. In our system the genetic algorithm is used to extract attribute from a huge attribute set.

2.2 K-means algorithm

K-means clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The method follows a simple way to cluster a given data set through a certain number of clusters fixed apriority. The main idea is to define k centers is one for each cluster. These centers should be arranged in a smart way because of different location causes different result in the clustering. So, the better choice is to arrange them as much as possible far away from each other. The next stride is to take each point belongs to the given data set and associate it to the nearest centre. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary centre of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centre. A loop has been generated, as a result of this loop it is notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more

2.3 MAFIA algorithm

MAFIA algorithm is used for mining maximal frequent item sets from a database. This algorithm is notably efficient when the item sets in the database are very large. The search procedure of the algorithm incorporates a depth-first traversal of the item set frame with effective pruning mechanisms.

Algorithm 1: K-means clustering Input:
The number of clusters k, and a database containing n objects. Output: A set of k clusters which minimizes the squared-error criterion. Method:
1) Arbitrarily choose k objects as the initial cluster centers;
2) Repeat
3) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
5) Until no change;

2.4 Decision algorithm

A decision tree is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The topmost node in a tree is the root node. Internal nodes are represented by rectangles, and leaf nodes are denoted by ovals. For the classification of an unknown sample, the sample attribute values are tested across the decision tree. A path is drawn from the root to a leaf node which holds the class prediction for that sample, so decision trees can easily be converted to classification rules

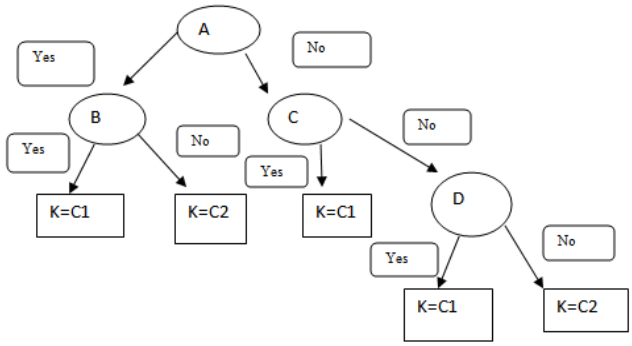

Fig.1. Representation of decision tree [1]

3. PROPOSED SYSTEM

In this project, we suggest by RNN Classifier as neural network ensemble that can incorporate different base classifiers into classifier ensembles models for classification problems. This project suggest that the impact of using different base classifiers on classification accuracy of RNN classifier ensemble. Classifier ensembles with five base classifier have used on five medical data sets. These results evaluated and compared choosing different type of decision tree algorithms for base classifier. The reliability of classification for most of datasets and classifier ensembles is increased when we select the appropriate RNN classifier achieves the minimum time required to build models. It is simple to understand and interpret and able to handle both numerical and categorical data, which requires little data preparation, for possible to validate a model using statistical tests, performs well with large datasets. It is robust, which means that performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

In proposed system using 5 datasets.
Classifier ensembles with five base classifier has used on five medical data sets,

they are

1. Transript : sample: walking speed,exercise.
2. Patients: sample: age,gender,weight,height.
3. Physical: sample: smoking status,
4. Diagnosis: sample: BP ,diabetics,cholesterol
5. Medication:Bmi,waist circumference.

3.1 MODULES DESCRIPTION

i. Pre-processing

Data pre-processing is an important step in the data mining process. It is process of data gathering methods are often loosely controlled and analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

ii. Cosmic Diabetics

Dataset Pre-processing Census Dataset is the procedure of systematically acquiring and recording information about the members of a given population. It is a regularly occurring and official count of a particular population. The term is used mostly in connection with national population and housing censuses, other common censuses include agriculture, business, and traffic censuses. It recommends that population censuses be taken at least every 10 years to cover census topics to be collected, official definitions, classifications and other useful information to co-ordinate international practice. COSMIC is an online database somatically acquired mutations found in human Diabetics. Somatic mutations are those that occur in non-germ cells that are not inherited by children. COSMIC, an acronym of Catalogue Of Somatic Mutations In Diabetics, curates data from papers in the scientific literature and large scale experimental screens from the Cancer Genome Project at the Sanger Institute.

iii. Data Cleaning

Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data pre-processing step while preparing the data for a data warehouse.

iv. Clustering Coefficients of Variation (CCV)

CCV is based on a very simple principle of variance-basis that finds a subset of features useful for optimally balancing the classification model induction between generalization and over fitting. CCV is founded on a basic belief that a good attribute in a training dataset should have its data vary sufficiently wide across a range of values, so that it is significant to characterize a useful model.

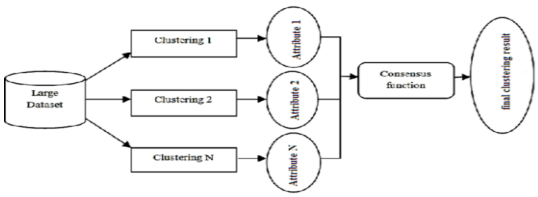

Fig 1. Clustering Process

3.2 Artificial Neural Networks

Artificial Neural Networks is the form of Neural Networks. A biological oriented network is formed to predict the diabetics. The representation of the neural network is given as:

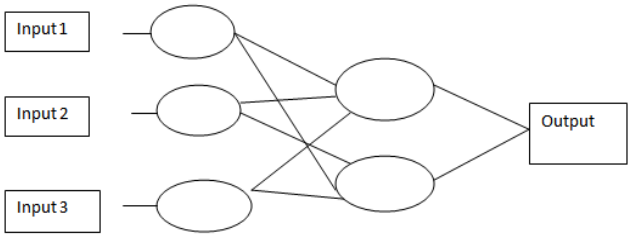

Fig 2. Representation of neural network [2]

ANN ought to be modified for every application else it leads to poor performance. The black box technology is the demerit discovered in ANN. Artificial Neural Networks (ANNs) has been utilized for the breast malignancy prediction in [19], [10]. In [19] the authors applied ANN on two distinctive breast malignancy dataset. Both of these datasets utilizes the morph metric attributes. An enhanced ANN model [30] has been utilized. Back propagation has been utilized to prepare the systems. Back propagation comprises of three layers a) Input layer b) hidden layer c) output layer. The predetermined model uses the likelihood nature and distinct the patients with the remarks good or bad. Several researchers utilized three-layer feedforwad ANNs with sigmoid function. In [1] used the ANN with multilayer perception's. It worked similar to the [19]. In any case, the distinction in the middle of [19] and [10] is the dataset. [10] Utilizes much greater dataset also, furnishes a decent investigation with the customary TNM neural framework. They guarantee that ANN can give vastly improved exactness when contrasted with the conventional TNM framework. ANN's accuracy achieved as high as 81%. It can even reach to 87% with the expansion of couple of more demographic variables.

3.3 RNN Classification

In general, RNN(Recurrent Neural Network) is decision tree based neural network are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. RNN correct for decision trees' habit of over fitting to their training set

RNN is a statistical algorithm that is used to cluster points of data in functional groups. When the data set is large and/or there are many variables it becomes difficult to cluster the data because not all variables can be taken into account, therefore the algorithm can also give a certain chance that a data point belongs in a certain group. This is how the clustering takes place. The algorithm clusters the data(Stages 1,2,3 and 4) in groups and subgroups. If you would draw lines between the data points in a subgroup, and lines that connect subgroups into group etc. the structure would look somewhat like a tree. This is called a tree. At each split or node in this cluster/tree variables are chosen at random by the program to judge whether data points have a close relationship or not.

The program makes multiple trees a.k.a. a forest. Each tree is different because for each split in a

tree, variables are chosen at random. Then the rest of the dataset (not the training set) is used to predict which tree in the forests makes the best classification of the data points (in the dataset the right classification is known).The tree with the most predictive power is shown as output by the algorithm.RNN Classifier algorithms are dynamic heuristic techniques. A RNN Classifier fitness function is calculated to find the RNN Classifier approach. In utilized the RNN Classifier systems for the forecast of breast tumour. This system is hybrid with the decision tree, ANN and logistic regression. They used 699 records acquired from the diabetics us patients at the University of Wisconsin. They utilized 9 indicator variables and 1 result variable for the information investigation with 10-fold cross approval. The researchers asserted that their RNN Classifier prediction model gives precision as much as 99%. Lipo Wang, Feng Chu, et al proposed the cancer prediction using gene expression data. They found the minimum gene probability. Two approaches were proposed namely, gene selection and gene ranking scheme. Based on the ranking scores, the prediction of the malignant diabetics is detected. They also employed T-test and class reparability. In a semi supervised ellipsoid method was proposed to detect the multiclass cancer classification. Each attribute is labeled and similar labeled data are clustered to detect the disease. The examining methods included artificial resembling of the information set. This can have the capacity to under examining the majority class and over examining the minority class or by consolidated the over and under sampling systems in an ordered way. An equalized data distribution is maintained. The oversampling method avoids the duplication of the data. The class imbalance problem is solved by the binary feature selection. The mean between two classes with its threshold is estimated to solve the duplication of the data. The feature values are binaries before setting the threshold value. Sometimes, there may be continuous feature selection in handling machine learning algorithms. These metrics method classified into Pearson correlation coefficient, feature assessment by sliding thresholds, FAIR and signal to noise correlation coefficient. These methods are designed to operate a continuous data and do not require any pre-processing of the data to work. All the experiments has conducted using MATLAB SPIDER package. Xiao et al distinguishing Differentially Expressed (DE) qualities from high-throughput gene expression estimations, by considering the parameters p- value and biological relevance. A better gene ranking was framed using Gene Set Enrichment Analysis (GSEA). The gene expression information acquired through such advances can be valuable for some applications in bioinformatics, if legitimately examined. For example, they can be utilized to encourage gene prediction. Several gene expression schemes do not provide the accuracy and efficiency of data mining systems. One of the primary difficulties in characterizing gene expression information is that the number of gene is normally much higher than the quantity of analyzed examples. Likewise, it is not clear which qualities are critical and which can be overlooked without lessening the classification improvement. Numerous example characterization methods have been utilized to break down microarray information. The characterization of the recorded examples can be utilized to arrange diverse sorts of dangerous tissues as in, where distinctive sorts of leukemia are distinguished, or to recognize harmful tissue from ordinary tissue, as done in, where tumour and typical colon tissues are investigated.

In this proposed, RNN algorithm is used to predicate heart disease.

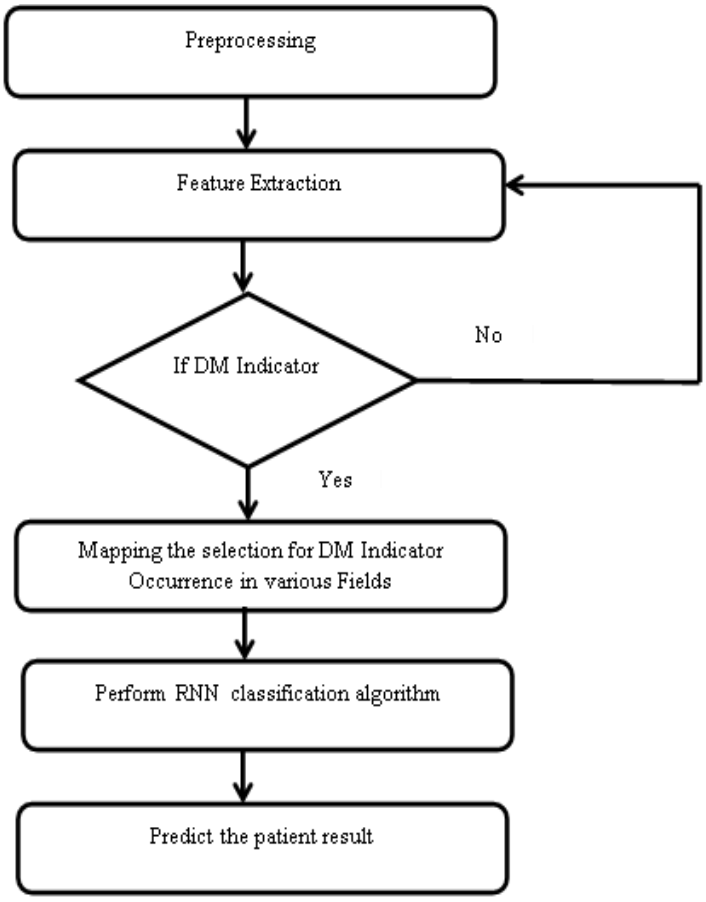

Fig 3.Proposed System

4. ACCURACY TABLE

In this table used to compare the accuracy between genetic algorithm and RNN algorithm. RNN algorithm consists more accuracy then other algorithm. The Proposed Method using RNN algorithm(0.98% - accuracy)

METHOD	ACCURACY	SPECIALITY	SENSITIVITY
RNN(RECURRENT NEURAL NETWORK)	0.98	0.9808	0.9847
GENETIC ALGORITHM	0.8366	0.8641	0.8043
K-MEAN	0.81	0.8395	0.777

Table1. Accuracy Table

4.1 Evaluation Metrics

By finding the confusion matrix these are parameters to find it:

Accuracy: the proportion of the total number of predictions that was correct.
Positive Predictive Value or Precision: the proportion of positive cases that were correctly identified.
Negative Predictive Value: the proportion of negative cases that were correctly identified.
Sensitivity or Recall: the proportion of actual positive cases which are correctly identified.
Specificity: the proportion of actual negative cases which are correctly identified.

5. CONCLUSION AND FUTURE ENHANCEMENTS

The data mining has played in an important role in healthcare industry, especially in predicting various types of diseases. The diagnosis is widely being used in predicting diseases; they are extensively used in medical diagnosing. In conclusion, there is no one data mining method to resolve the issues in the healthcare data sets. In order to obtain the highest accuracy among classifiers which is important in medical diagnosing with the characteristics of data being taken care, we need to design a hybrid model which could resolve the mentioned issues. In future more work will be done in this field so to improvise the treatments and the lifetime of patient by properly maintaining and analyzing the health sector data for directions is to enhance the predictions using hybrid models.

REFERENCES

1) Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, Hena M" Heart Disease Diagnosis Using Data Mining Technique"-2017

2) Marc E. Lippman, "Harrison's Principles of Internal Medicine", 16th ed., Ch. 76, "Diabetics," by World Health Organization "Fact sheet No. 297: Cancer".

3) Umer Khan, Hyunjung Shin, Jong Pill Choi, Minkoo Kim, "wFDT - Weighted Fuzzy Decision Trees for Prognosis of Diabetics Survivability", AusDM 2008.

4) Dursun Delen, Glenn Walker, Amit Kadam, "Predicting diabetics survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, Volume 34, Issue 2, Pages 113-127, June 2005.

5) Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin and Minkoo Kim, "Predicting Diabetics Survivability Using Fuzzy Decision Trees for Personalized Healthcare", EMBS 2008.

6) J. Bundred, "Prognostic and predictive factors in diabetics", Cancer Treatment Reviews, Vol. 27, Issue 3, Pages 137-142, June 2001.

7) Yijun Sun, Steve Goodison, Jian Li, Li Liu and William Farmerie, "Improved diabetics prognosis through the combination of clinical and RNN Classifier markers", Bioinformatics 2007.

8) Wei-Pin, Chang, Der-Ming, Liou "Comparison of Three Data Mining Techniques with RNN Classifier Algorithm in the Analysis of Diabetics Data", Journal of Telemedicine and Telecare, 2008, 9.

9) J.R, QUINLAN, "Induction of Decision Trees", Journal of Machine Learning, Volume 1, Number 1, March, 1986.

10) David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams, "Learning representations by back-propagating errors", Letter to Nature, 1986.

11) Burges C, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, 1998.

12) Chih-Lin Chi, W. Nick Street, William H. Wolberg, "Application of Artificial Neural Network-Based Survival Analysis on Two Diabetics Datasets", AMIA Annu Symp Proc. January, 2007.

13) Street, W. N. "A neural network model for prognostic prediction" Fifteenth International Conference on Machine Learning", pages 540-546, San Francisco, 1998.

14) Harry B. Burke, Philip H. Goodman, David B. Rosen, Donald E. Henson, John N. Weinstein, Frank E. Harrell, Jr., Jeffrey R. Marks, David P. Winchester, David G. Bostwick, "Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction", Cancer, Feb. 1997.

15) S. B. Kotsiantis, I. D. Zaharakis, P. E. Pintelas, "Machine learning: a review of classification and combining techniques", Artificial Intelligence Review, 2007.

16) Jaree Thongkam, Guandong Xu and Yanchun Zhang, Fuchun Huang, "Toward diabetics survivability prediction models through improving training space", Expert Systems with Applications, Volume 36, December, 2009.

17) Y.-J. Lee, O. L. Mangasarian, and W. H. Wolberg, "Diabetics Survival and Chemotherapy: A Support Vector Machine Analysis", Data Mining Institute, Computer Sciences Department, University of Wisconsin, 2000.

18) Krzysztof Fujarewicz, Malgorzata Wiench, "Selecting differentially expressed genes for colon tumor classification" int.j.Appl.Math.Comput.Sci, 2003. Vol.3, No.3, Pg.no:327-335

19) D. Lavanya, Dr.K.Usha Rani, "Performance Evaluation of Decision Tree Classifiers on Medical Datasets",. International Journal of Computer Applications 26(4):1-4, July 2011

20) Osmar R. Zaïane, Principles of Knowledge Discovery in Databases. [Online]. Available webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/ch1.pdf

21) Street W.N., "A Neural Network Model for Prognostic Prediction", Fifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann, 1998.

22) Sujatha, Dr.K.Usha Rani, "Evaluation of Decision Tree Classifiers on Tumor Data sets", IJETTCS, Vol2, Issue4, July-aug2013, Pg.no:418-423

23) Sujatha, Dr.K.Usha Rani,"An Experimental Study on Ensemble of Decision Tree Classifiers", IJAIEM, Vol 2, Issue 8,August 2013,Pg.no:300-306

24) Val´erie Bourd`es, St´ephane Bonnevay, Paolo Lisboa, R´emy Defrance, David P´erol, Sylvie Chabaud, Thomas Bachelot, Th´er`ese Gargi,6 and Sylvie N´egrier "Comparison of Artificial Neural Network with Logistic regression as Classification Models for Variable Selection for Prediction of Diabetics Patient outcomes"

25) I Guyon, J Weston, S Barnhill "Gene selection for cancer classification using support vector machines" .Machine learning, 2002 – Springer

26) Z. Chen, "Research of Data Mining Based on Neural Network", E-Product E-Service and Entertainment (ICEEE), 2010 International Conference on: IEEE, (2010), pp. 1-3.

27) G. K. Dhondalay, C. Lemetre and G. R. Ball, "Modeling estrogen receptor pathways in diabetics using an Artificial Neural Networks based inference approach", Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS International Conference on: IEEE, (2012), pp. 948-951.