Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 4, April 2018, pg.183 – 190

Anticipating Thyroid Disorders using Data Mining Techniques

S.Sangeetha[1], K.Palanivel[2]
[1]Research Scholar, [2]Associate Professor
Dept. of Computer Science, A.V.C. College (Autonomous), Mayiladuthurai – 609 305, TN, India
[1] sangeetha2achieve@gmail.com

Abstract— Anticipating the Thyroid disease is one of the current focuses in Medical research. The most difficult undertaking of restorative field is to give the infection analysis at the beginning period with higher precision. The sickness forecast assumes a critical part in information mining. Information mining is mostly utilized as a part of medicinal service associations for basic leadership, diagnosing malady and giving better treatment to the patients. It is a procedure of investigating huge informational collections to discover a few examples. These examples can be useful for forecast demonstrating. In this theory, the principle objective is an order of thyroid ailment. An investigation of thyroid sickness finding systems utilized as a part of information mining. Thyroid is one of the impulsive syndromes in restorative field. Thyroid hormones control the body's metabolic rate. Also, information mining methods have been connected in different areas the order, after effects of the restorative informational index which helps the method for medications to the patients. In proposing procedure, look at different calculations, for example, naval forces' bayer, random backwoods, irregular tree, REP Tree, D tree to discover which calculation to give the best outcome for thyroid ailment expectation. An exploratory examination is conveyed out our calculation to accomplish better exactness. There are numerous information mining grouping Algorithms, for example, naval forces, REP Tree, arbitrary tree, irregular forest, D tree and k-overlap cross approval et cetera. The proposed Algorithm gives exactness 99.80% with cross approval k=6.

Keywords— Classification, D tree, Information mining, REP Tree, Thyroid disease, WEKA.

I. INTRODUCTION

Medicinal finding can be seen as an example grouping issue: based an arrangement of information includes the objective is to characterize a patient as having a specific issue or as not having it. Thyroid hormone issues are the most predominant issues these days. In this proposal a simulated neural system approach is created utilizing a back proliferation calculation with a specific end goal to analyse thyroid issues. It gets various factors as info and produces a yield which gives the after effect of whether a man has the issue or is solid. It is discovered that back proliferation calculation is ended up being having high affectability and specificity. Thyroid is a butterfly molded organ discovered just underneath the Adam's apple of our neck. It is in charge of the digestion exercises of our body. When it works legitimately it produces two hormones called triodiothrinine (T3) and thyroxin (T4). A hormone called Thyroid Stimulating Hormone (TSH) which is emitted by pituitary organ is additionally in charge of T3 and T4 hormones. TSH, T3 and T4 chooses the wellbeing of a man. Over action of thyroid hormones brings about hyperthyroidism where us the under movement of similar outcomes in hypothyroidism.[9]

II. DATA MINING

Information mining is the figuring procedure of finding designs in substantial informational indexes including strategies at the crossing point of manmade brainpower, machine learning, insights, and database frameworks. It is an interdisciplinary subfield of software engineering. The general objective of the information mining process is to remove data from an informational collection and change it into a reasonable structure for additionally utilize. Beside the crude investigation step, it includes database and information administration perspectives, information pre-preparing, model and surmising contemplations, intriguing quality measurements, unpredictability contemplations, post-handling of found structures, representation, and web based refreshing. Information mining is the investigation venture of the "learning disclosure in databases" process, or KDD.

Information mining, the extraction of concealed prescient data from expansive databases, is an effective new innovation with extraordinary potential to enable organizations to focus around the most essential data in their information stockrooms. Information mining instruments foresee future patterns and practices, enabling organizations to make proactive, learning driven choices.[8] The computerized, planned examinations offered by information mining move past the investigations of past occasions gave by review apparatuses average of choice emotionally supportive networks. Information mining devices can answer business addresses that customarily were excessively tedious, making it impossible to determine. They scour databases for concealed examples, finding prescient data that specialists may miss since it lies outside their desires. Most organizations effectively gather and refine monstrous amounts of information. Information mining strategies can be executed quickly on existing programming and equipment stages to improve the benefit of existing data assets, and can be coordinated with new items and frameworks as they are expedited line.

III. KNOWLEDGE DISCOVERY

Learning disclosure is a procedure that concentrates certain, conceivably valuable or already obscure data from the information. The information disclosure process is portrayed as takes after:

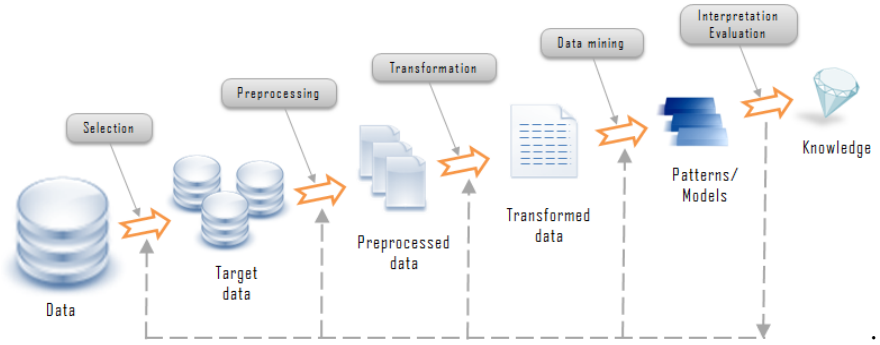

Figure 1. Steps for Knowledge discovery in Database[1]

- Data originates from assortment of sources is incorporated into a solitary information store called target information.
- Data at that point is pre-handled and changed into the standard arrangement.
- The information mining calculations process the information to the yield as examples.
- Then those examples and standards are deciphered to new or valuable learning or data.

IV. THYROID DISEASE

Thyroid disorders are conditions that affect the thyroid gland a butterfly-shaped gland in the front of the neck. The thyroidhas important roles to regulate numerous metabolic processes throughout the body. Different types of thyroid disorders affect either in structure of function. Some specific kinds of thyroid disorders are: A) Hypothyroidism; B) Hyperthyroidism.[15]

A) HYPOTHYROIDISM

Hypothyroidism results from the thyroid gland producing an insufficient amount of thyroid hormone. It can develop from problems within the thyroid gland, pituitary gland or hypothalamus.

B) HYPERTHYROIDISM

Hyperthyroidism describes excessive production of thyroid hormone, a less common condition than hypothyroidism. Symptoms of hypothyroidism usually relate to increased metabolism.

V. LITERATURE SURVEY

(Roshan Banu and Sharmili, 2017) performed a research on predicting the Thyroid disease in advance. Thyroid disease is across the board around the world. The thyroid gland is an organ at the lower section of the human neck that produces the hormones that helps to regulate many process like growth, energy balance, body temperature and heart rate which is in charge of delivering wide-range of thyroid hormones. Data mining technique is mainly used in healthcare organizations for decision making, diagnosing disease and giving better treatment to the patients. Classification of this thyroid disease is an important task. The algorithms used in this thesis are CART, LDA, classification, clustering, decision tree and k-fold cross validation.

Most tedious and challenging task is to provide disease diagnosis at early stage with higher accuracy in the medical science field. The disease prediction plays an important role in data mining. Data mining is a process of analyzing and extracting hidden information from large data sets to find some patterns. These patterns are useful in prediction method. Clinics and hospitals collect a large amount of patient data over the years. These data provides a basis for the analysis of risk factors for many diseases. There are various types of diseases predicted in data mining namely lung cancer, liver disorder, breast cancer, thyroid disease, diabetics etc.[14] Predicting thyroid disease is analyzed in this paper. Thyroid gland will stow thyroid hormones to maintain the body's metabolic rate. Thyroid disorders are caused due to the malfunction of thyroid hormones. Thyroid or thyroid gland releases diiodothyronine (T3) and thyroxin (T4) into the blood stream as the vital hormones. Thyroid hormones functions are to regulate the rate of metabolism and effect the growth. There are two most common problems of the thyroid disorder or thyroid disease they are hyperthyroidism and hypothyroidism. Hyperthyroidism releases too much thyroid hormone into the blood due to over active of thyroid. This can stimulate your body's metabolism significantly, symptoms like Table 1 and Table2. Hypothyroidism is when the thyroid is not active and releases too low thyroid hormone into the blood.[7] This upsets the normal balance of chemical reactions in your body. It seldom causes symptoms in the early stages, but, over time, untreated hypothyroidism can cause a number of health problems, such as obesity, joint pain, infertility and heart disease.

The process carried out in this is the original sample is randomly partitioned into k equal size subsamples, among the k sample a simple is retained as the validation data for testing the model, now the remaining k-1 subsamples are used as training data. This cross validation process is carried out k folds with each of the k samples used one time of the data to validate. The obtained k results are combined to produce a single estimation. The dominance considered in this method over repeated random sub-sampling is that all observations observed are used for both training and validation, and each observation is used for validation exactly the data is loaded into weka software in this thesis they use this weka open source software to carry out their experiment)[3]. After pre-processing, various data mining classification techniques are applied to develop the predictive models on the data set. And then system is trained using the training set. After that it is trained and tested using up to 10 fold cross validation method. Evaluation is performed using certain performance measures. In this LDA gives better accuracy and is measured as 99.62 with cross validation k=6. The carried work using LDA algorithm has the main disadvantage that it has unbalanced design (i.e. the number of objects in various classes are highly different) it is also not applicable for non-linear problems.

It also helps to divide the data in dataset according to described disorders. This method provides five different splitting conditions for the construction of decision tree. The conditions are Information Gain, Gain Ratio, Gini Index, Likelihood Ratio Chi-Squared Statistics, and Distance Measure[4]. Among, the above splitting conditions three belongs to Impurity based splitting condition and other two are Normalized Impurity based splitting criteria. As a result, the decision tree classifies the thyroid data-set into three classes of disorders. When the data-set is small, splitting criteria will make it. But for large data-set, it is difficult to produce more accurate decision tree. Various splitting rule for decision tree attribute selection had been analysed and compared. This helps to diagnosis the thyroid diseases through the extracted rules. From this experiment, it is clear, that normalized based splitting rules have high accuracy and sensitivity or true positive rate. This work can also be extended for any medical datasets. Further enhancement can be made by using various optimization algorithms or rule extraction algorithms.

VI. METHODOLOGY USED

The accompanying advances are incorporated into the characterization procedure of this paper. Three diverse classifier is picked Naïve Bayes, Random Forest, Random Tree, D Tree. WEKA device is utilized to break down the anticipated esteems by every one of the classifier. The Precision, Recall and F-Measure of every classifier is figured. At long last the outcome is investigated and the best execution calculation recognized.

NAIVE BAYES

The UCI dataset for the determination of hepatitis utilizing both regulated and unsupervised ANNs. We have utilized three kinds of systems out of which two have a place with the administered class; the Feedforward Backpropagation Neural Network (FFNN) and Generalized Regression Neural Network (GRNN) and one has a place with the unsupervised sort; Self Organizing Map (SOM). A performance correlation has additionally been appeared between the networks we utilized and the outcomes were likewise contrasted and the past investigations that utilized the same dataset for the conclusion and order of hepatitis. In this stage the information is first gathered. Amid the gathering the characteristics have been talked about with the master specialist and the objectives (accessible as classes in the UCI dataset) of each preparation test are set. In this manner, the whole informational index is separated into the preparation and testing sets. In the event of the approval set the preparation set is additionally isolated into two sections; one for preparing and the other for approval. From that point, this information is prepared to be exhibited to the ANNs. It involves 155 examples which incorporates both the positive and negative cases for hepatitis infection.[6] All examples have 19 characteristics, which are exhibited in table I. Out of these 19 properties 13 have parallel esteems (it is possible that it is No or Yes) and the rest 6 have a scope of qualities. These 19 qualities incorporate 12 physical examination tests, 5 Liver Function Tests (LFT's) and 2 general traits of a patient.

RANDOM TREE

The human master plays out the conclusion of skin infections by gathering quiet records and dissensions. This rundown of patient grumblings and watched skin conditions is then ventured into a few Boolean side effects. The indications are additionally subjected to learning coordinating with the information effectively controlled by the human master (learning base-involvement). In the event that there is a match, the specialist prescribes the sickness as a conceivable skin malady. Sometimes, the human master may subject the patient to facilitate research facility tests all together to ascertain the causative specialist of the skin condition. The test could fill in as a corroborative test if the malady analyzed is really caused by microorganism, for example, microbes, warts, infection, organisms, etc. When the human master is unpracticed or has not gone over such skin condition, he utilizes experimentation to analyse[5]. This is finished by the blend of all the conceivable conditions, contrasting them and known conditions and narrowing the judgment. Amid this procedure, learning is said to have occurred if the skin condition is appropriately analyzed and treated. In this manner, the human master depends to a great extent on his experience and the patient protestation translation.. Disease study incorporates the general accumulation and posting of the skin sicknesses that is considered and utilized as a part of the procedure of framework advancement. This incorporates grumbling, perception and lab test. The information utilized for the indicative framework comprise of the accompanying segments: Patient essential signs, Patient verbal protestations, Patient socioeconomics and Presence of particular side effects. The components of every one of the segments fill in as an info variable to the system. In this manner, the assignment of the counterfeit neural system is to draw a relationship between's simply the patient's introduction, utilizing revealed side effects and key sign. While quiet socioeconomics and indispensable signs are a key component in giving pieces of information to a sickness, it has been watched that the patients' grumblings give more prominent experiences to foreseeing restorative analyses.

RANDOM FOREST

The term originated from irregular choice woods that were first proposed by Tin Kam HO of Bell lab1s in 1995. Arbitrary timberland is an outfit classifier that comprises of numerous choice trees and yields the class that is the made of the classes yield by singular trees. It is given freely some controlled adjustment. Trees and the outcomes included indiscriminately woods depends on larger parts of precise yield. In dataset, where M is the aggregate number of information ascribes to the dataset, just m credits to picked indiscriminately for each tree where m<M [8].

A true endeavor to show, a customized system for medicinal services application for basic leadership. Through the analysis procedure with the utilization of fluffy choice tree calculation, the conclusion message is gotten. In view of the demonstrative message got, the SWRL rules are executed to create the treatment stream. The system comprises of three conditions, for example, Fuzzy lead age and Diagnosis, Rule execution and Ontology construction. Ontology spoke to in OWL design is utilized to build a learning base. OWL might be separated into three sub dialects I) Owl Full ii) Owl DL iii) Owl Lite. The three classifications are shaped based on expressiveness. The proposed think about utilizations the Owl DL to build the sustenance piece cosmology. To start with the help of a dietician, we gather and investigate the ideas and ascribes as for the nourishment estimations of different sustenance things. Three classes portrayed in meta-metaphysics are spoken to as nourishment classifications, determination result and patient profile data. For instance nourishment classifications incorporate six gatherings of sub classes: Grains and Starches, Fruits and Juices, Vegetables, Pulses, Fat or Oils and Meats. The conclusion result class has nine ideas: the initial three ideas for thyroid organ (hyperthyroid, hypothyroid and typical) and the staying six ideas for corpulence administration. Understanding data class incorporates the insights about patient's clinical and individual data. It shows the halfway outline of Food Composition Ontology (FCO) for thyroid organ administration in which the finding message class has three people: hyperthyroid, hypothyroid, and typical.[4]

D TREE

Progressive various classifier order plot, which protects the quality of the different classifier approach and furthermore figures out how to diminish a portion of the issues looked by other numerous classifier calculations. In our plan, the framework asset prerequisites are decreased as is the preparation time. From the outcomes on pap-spread information, it can be seen that our approach delivers preferable execution over other numerous classifier calculations and superior to anything a classifier created by human specialists. Calculation is utilized for anticipating the thyroid ailment with the related indications. The proposed calculation fills in as takes after. At first, all articles in the informational index are thought to be unassigned. At that point picks a subjective unassigned question p from the informational collection. Additionally order the dataset utilizing Hierarchical different classifier grouping plan, which saves the quality of the numerous classifier approach and furthermore figures out how to lessen a portion of the issues looked by other various classifier calculations. Along these lines the information are arranged in effective way give precise data. The client can anticipate and test their wellbeing with the side effects. The client can foresee the thyroid ailment with related side effects.[1]

VII. STATISTICAL MEASURE

The exactness of the classifier is given by TP rate, FP Rate, Recall, F-Measure, Precision utilizing WEKA device. WEKA instrument is an effective programming stage that gives a coordinated situation for machine learning, information learning, content learning and different business and forecast investigation.

The genuine truth about the classifiers is given by false positive rate, genuine positive rate, review, accuracy and F-measures utilizing WEKA instrument. WEKA is an effective programming proclamation of conviction that gives a comprehensive domain for hardware learning, information mining, content mining and other exchange and forecast examination. The not too bad of measures from by and large advised the classes has been taken to attempt the amid measure for classifiers. For solid outline, to attempt the around exactness for a classifier for a subject to dataset, satisfactory of precisions of both (genuine/false) classes is figured.

1) Precision

Accuracy is the accuracy or precision of roughly arranged class, legitimately known as positive prescient esteem. It is the symmetry of occasions which initially have class x/Total delegated class x. So essentially important exactness coordinated the precise outcomes and it takes on the whole applicable information yet returns just highest outcomes. In rapidly, it is the quantity of chose things which were connected.

2) Recall

Accuracy = (True Positive/(True Positive + False Positive))*100;

Review to give affectability of issue and it forms esteems or item amount or fulfilment. It restored the most important and part from the records that are applicable as result from the question. As it were modules that are extremely perceived as hard to keep up from the aggregate number of modules. To put it plainly, it is the quantity of related articles that were picked.

Review = (True Positive/(True Positive + False Negative))*100;

*** True Positive**

Genuine positive are the positive tuples which were suitably named each classifier. It is the extent arranged as class x/Actual aggregate in class x. Genuine positive anticipated separately modules that are anticipated really as the outcomes determined toward the end.

Genuine Positive rate = (True Positive/(True Positive + False Negative))*100;

*** False Positive**

False positive, extent erroneously ordered as class x/Actual aggregate of all classes, with the exception of x. It is mistakenly anticipated contrasted with unique outcomes.

False Positive rate = (False Positive/(False Positive + True Negative))*100;

3) F-Measure

F-Measure sorted as (2*Precision*Recall/(Precision+ Recall))*100. It is a joined measure for accuracy and review.

VIII. EXPERIMENTAL RESULTS

In the trials, Navies gives the performance outcome with throughput of 50%. RF gives an outcome with throughput of 88%. Random tree gives an outcome with throughput of 67%. DT gives an outcome with throughput of 90%.

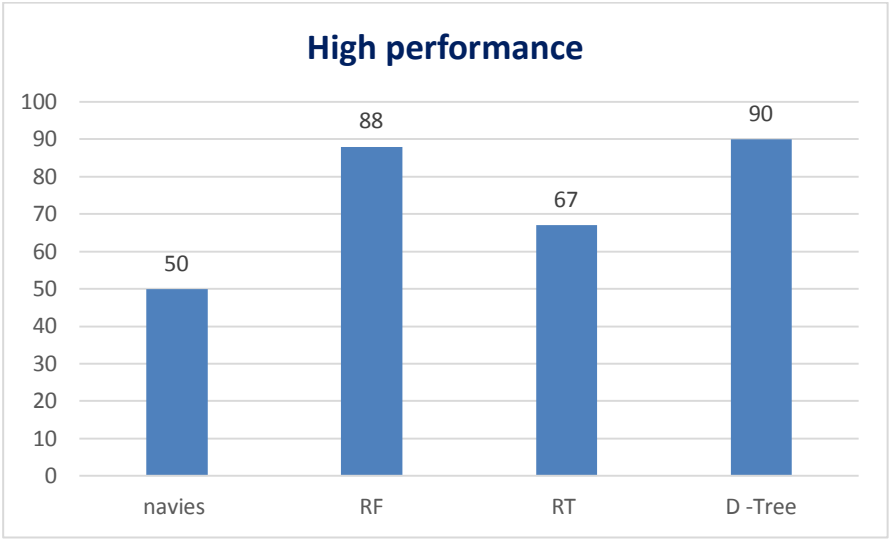

Figure 2. High Performance Diagram for comparing time, throughput, TP, FP with different algorithms

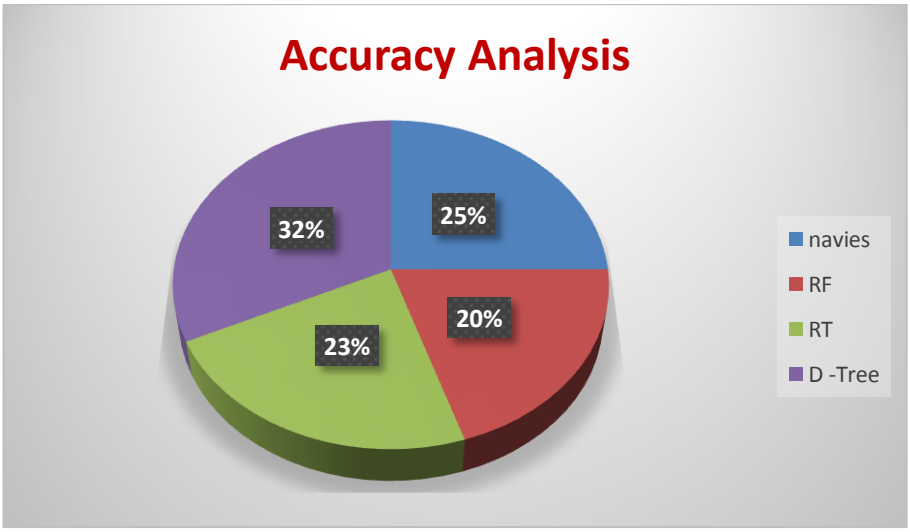

Figure 3. Algorithms performance based on Accuracy

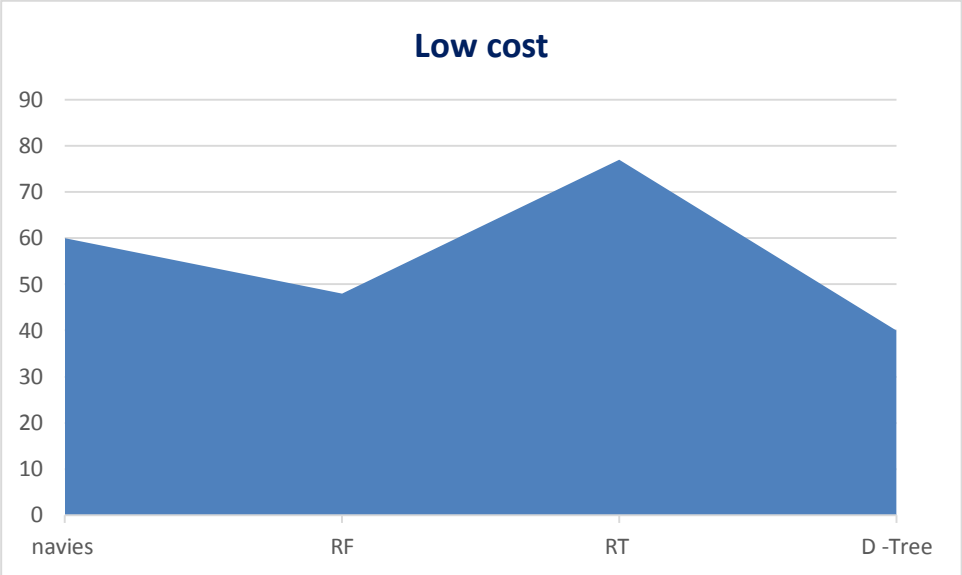

Figure 4. Low cost when compare with different algorithms

IX. CONCLUSION AND FUTURE WORK

Thyroid sickness is one noteworthy malady and expectation is the exceptionally troublesome assignment. This model gives with grouping and bunching precision with minimum number of highlights contrasted with other the current created display. Different part lead for D tree quality choice had been broke down and thought about. In this manner the review distinguishes the how information mining methods to anticipate the thyroid issue at a prior stage. Diverse Researchers have proposed distinctive procedures to anticipate the thyroid issue and various types of precision level according to utilized methods. These strategies help to limit the clamour information of the patient's information from the information bases. In this postulation, they have connected D tree data mining characterization methods is utilized to arrange the thyroid ailment. K-overlap cross approval is additionally performed. The D tree Algorithm gives 99.80% exactness with k=6 folds cross approval. In future, a superior strategy to analyse thyroid issue can be discovered with changes in the proposed techniques. What's more, a similar method is utilized to apply for other malady datasets, for example, coronary illness, diabetes et cetera.

REFERENCES

[1] P. Yasodha , M. Kannan M, "Analysis of Population of Diabetic Patient Database in WEKA Tool", International Journal of Science and Engineering Research, VoL.2 Issue.5, 2011.

[2] S. Vijayarani , S. Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease Prediction", International Journal of Innovative Research in Computer and Communication Engineering, Vol.1, Issue.3, pp.735-741, 2013.

[3] Dr.V.Karthikeyani , I.Parvin Begum "Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease" International Journal on Computer Science and Engineering (IJCSE) Vol. 5 No. 03 Mar 2013 205-210

[4] S. Tirunagari, N. Poh, K.Aliabadi, D.Windridge & D.Cooke, "Patient level analytics using self-organising maps: A case study

[5] Dr. G. Rasitha Banu, Baviya "A study on Thyroiddisease using Data Mining Technique", IJTRA Journal,aug 2015.

[6] Dr .G .Rasitha Banu, Baviya "predicting Thyroiddisease using Data Mining Technique ", IJMTER journal, March 2015.

[7] Pandey, Rohit Miri , Tandan "Diagnosis AndClassification Of Hypothyroid Disease Using DataMining", International Journal of Engineering Research& Technology (IJERT),Vol. 2 Issue 6, June – 2013

[8] D .Lavanya, Dr .Usha Rani, " Performance Evaluation ofD tree Classifiers on Medical Datasets", International Journal of Computer Applications (0975 –8887),Volume 26– No.4, July 2011

[9] K. Saravana Kumar, Dr. R. Manicka Chezian "SupportVector Machine And K- Nearest Neighbor BasedAnalysis For The Prediction Of Hypothyroid",International Journal of Pharma and BioSciences.,0ct.2014.www.ijpbs.net
[10] Prerana, Parveen Sehgal, and Khushboo Taneja, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", International Journal of Research in Management, Science & Technology (E-ISSN: 2321-3264) Vol. 3, No. 2, April 2015.
[11] K. Rajesh , V. Sangeetha, "Application of data mining methods and techniques for diabetes diagnosis" . International Journal of Engineering and Innovative Technology (IJEIT), Vol.2,Issue.3, pp.224.
[12] I.Parvin begum, V. Karthikeyini, K. Tajuddin, I. Shahina Begum, "Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction", International journal of Computer Applications, Vol.60, Issue.12, pp. 26-31, 2012.
[13] P.P.Dhakate, S. Patil, K. Rajeswari, D.Abin "Preprocessing and Classification in WEKA Using Different Classifier", International Journal of Engineering Research and Applications, Vol.4, Issue.8, pp.91-93, 2014.
[14] Dr. V. Karthikeyini, I. Parvin Begum," Comparison a Performance of Data Mining Algorithms (CPDMA) in Prediction Of Diabetes Disease", International Journal on Computer Science and Engineering (IJCSE), Vol.5 Issue.3, 2013.
[15] A. Rajput, R.P.Aharwal, M. Dubey, S.P. saxena "J48 and JRIP Rules for E-Governance Data" International Journal of Computer Science and Security, Vol.5, Issue.2, pp.201-207, 2011.

