IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE STUDY OF MACHINE LEARNING BASED MODEL FOR THYROID DISEASE PREDICTION

[1]SURESH KUMAR KASHYAP, [2]DR. NEELAM SAHU

[1]Research Scholar Ph. D(CS), Dept. of IT & CS, Dr.C.V.Raman University, Kota,Bilaspur (CG) India
[2]Associate Professor Dept. of IT & CS, Dr. C.V. Raman University, Kota,Bilaspur (CG) India

***ABSTRACT*:** Thyroid disease is very common, with an estimated 20 million people in the United States having some form of thyroid disease and research studies emphasizing that approximately 43 million people in India suffer from thyroid disease. The thyroid gland is a small organ found earlier in the neck. It is shaped like a butterfly. Thyroid problems is a abnormal production of thyroid hormones. This research paper based on the prediction of thyroid disease using Decision Tree and Artificial Neural Network. This paper explains to see the separation of the best way. As a result, the operation will be completed in both segregation modes and their accuracy will be compared with the confusion matrix. It has been concluded that ANN provides better accuracy than other decision tree classifier.

KEYWORDS: **Classification, Thyroid, Hyper Thyroid, Hypo Thyroid, Confusion matrix.**

I. INTRODUCTION

According to a study by "DAILY TIME OF INDIA" when one in ten people in India suffers from thyroid disease. The thyroid is in ninth place compared to other common types of disease such as diabetes, allergies, asthma, cholesterol, depression, diabetes etc. The thyroid gland in addition is a hormone found in the lower part of the neck that produces hormones that help regulate many body processes, including growth, energy balance, body temperature, and heart rate. The endocrine releases thyroxine (T4) and triiodothyronine (T3) into the bloodstream because they are the main hormones. The functions of thyroid hormones are to regulate the speed of metabolism and affect growth. There are two most common complications of thyroid disease or thyroid disease. In Hyperthyroidism - it releases too much thyroid hormone due to thyroid function and Hypothyroidism - when the thyroid is inactive and releases very low hormone in the blood.

Machine learning is a branch of artificial intelligence .It is a way of automatically analyzing data to create an analytical model. It is based on the idea that programs can learn from data, identify patterns and make decisions with minimal human intervention.

II. LITERATURE SURVEY

Maximum of the people are not willing to spend time and money to know the prediction for thyroid disease. Banu et al system explains about people to know the prediction for thyroid disease and also to know the prediction details and level of disease anywhere in the world. They used classification method to find the prediction details. Iodine acquires part a noteworthy role of the thyroid organ. It empowers thyroid hormones, and is essential for their production. Iodine existed in food and water.

Sayyad Rasheeduddin et.al research work is based on unsupervised Graph Clustering Optimization based Extreme Machine Learning technique to detect threat of thyroid. This approach identified factors affecting risk of thyroid disease. Author found approach superior than univariate logic. The ultrasound technique used for identification found popular, affordable and efficient .

Liyong Ma, Chengkuan Ma et.al research work is based on proficient way using convolutional neural network to detect disease illnesses on SPECT datasets. Suggested approach result worked better than existing methods .

K. Rajam et.al research work is based on data mining supervised functionalities Naïve bayes, decision tree, back propagation, Support vector machine identifies thyroid disease at former phase. Outcomes evaluated based on parameters speed, accuracy, performance and cost and found effective for treatment of the patient .

Marissa Lourdes De Ataide et.al work is based on applied a two multilayer perceptron classifier for classifying thyroid diseases into three classes as thyroid, hyperthyroid and hypothyroid and to classify hypothyroid disease into primary, secondary and tertiary hypothyroid with focused on maximum accuracy in minimum time. This classifier gave good accuracy of classification.To reduce problem applied neural network for analyzing thyroid problem.
Fatemeh Saiti et.al, applied Genetic Algorithms Using Support Vector Machine for thyroid verdict . G. Rasitha Banu has predicted problem using Linear Discriminant Analysis (LDA)- Data Mining approach.

K. Vembandasamy, R. Sasipriya, E. Deepa (2015). "Aims on analyzing heart diseases using a Naïve Bayesian algorithm". The algorithm used here is Naïve Bayes, which firmly assumes that the presence of any attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent. The tools used are WEKA and classification is done by splitting data into 70% of the percentage split. The naïve Bayes technique used was able to produce 86.41% of the input data correctly and 13.58% of inaccurate instances. He uses a dataset collected from a leading diabetic research institute in Chennai which has about 500 instances or patients.

III. METHODOLOGY

The architectural process to perform classification techniques for prediction of diseases: To perform the classification process, the Input data is loaded using databases. Input Data can be with missing values, noisy values, in consistent data. Such data need to be pre-processed in a pre-processing step. For better classification, optimal selection of features plays a major step. Best features are selected and for further process of classification, the dataset is divided as training and testing data. Training data is loaded to the classification algorithm and validated using testing data. Overall performance of the data is evaluated through accuracy or recall or precision or speed etc. Finally, results show the prediction values. Figure 1 describes the flow diagram of the Prediction Process.

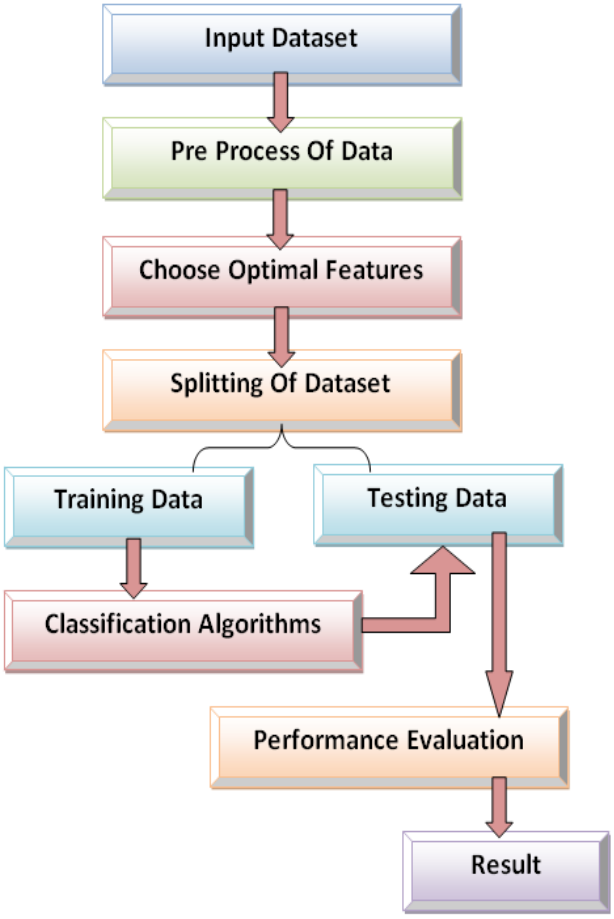

Figure 1:Architecture of Thyroid Prediction System.

Data Set Description:-The data set used for experimental purpose we can download from UCI machine learning Repository that was used for implementation with 3772 instances of 23 independent attribute and 1 dependent attribute. The details of data set is shown next slide.

Where class is varied as following

- Normal - 1
- Hyperthyroidism - 2
- Hypothyroidism – 3

Thyroid Disease data set Attribute description

S.NO.	Attribute Name	Value Type	S.NO.	Attribute Name	Value Type
1	Age	continuous,?.	13	goitre	f,t.
2	Sex	M,F,?.	14	TSH_measured	f,t.
3	on_thyroxine	f,t.	15	TSH- Thyroid Stimulating Hormone	continuous,?.
4	query_on_thyroxine	f,t.	16	T3_measured	f,t.
5	on_antithyroid_medication	f,t.	17	T3 - Total Triiodothyroxine	continuous,?.
6	thyroid_surgery	f,t.	18	TT4_measured	f,t.
7	query_hypothyroid	f,t.	19	TT4- Total Thyroxine	continuous,?.
8	query_hyperthyroid	f,t.	20	T4U_measured	f,t.
9	pregnant	f,t.	21	T4U	continuous,?.
10	sick	f,t.	22	FTI_measured	f,t.
11	tumor	f,t.	23	FTI – Free Thyroxine Index	continuous,?.
12	lithium	f,t.			

Table 1: Thyroid Disease data set Attribute description

Methodology section consists of three parts, that are following:-

A. DATA PREPROCESSING:

Pre-data processing (DP) is a very neglected but a major process. Data processing includes, cleaning, generalization, feature removal and selection etc. Raw data has a lower signal than audio, lost values, and inconsistencies that affect the results of data processing, Partitioning to help improve the data aspect, improve efficiency and simplicity in the mining process. Data pre-processing has been following categories:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

B. FEATURE SELECTION / ATTRIBUTE REDUCTION:-

It is in the feature selection methods to identify the most suitable components to be categorized and will be further categorized as sub-selection methods. Feature selection, is effective in reducing size, by removing unwanted and unimportant data, increasing learning accuracy, and improving understanding of results. Feature selection algorithms have two broad categories. They are:

A. The filter model- This model relies on standard training data features to select other features without including any learning algorithm. The filtering model examines the correlation of factors from data only, independent of classifications, using measures such as distance, information, dependence and consistency.
B. The wrapper model- This model requires a single learning algorithm pre-determined for feature selection and uses its functionality to determine which features are selected. for each of the new features produced, the folding model should be informed of the separator hypothesis. it is a tendency to hunt for features that are better suited to a predetermined learning algorithm that lead to higher learning performance, but also often require more calculation time and are more expensive than the filtering model.

C. MODEL EVALUATION:-

Model evaluation is an important part of any model development process. It helps to find the most effective model that represents our data and how the chosen model will add in the long run. it is a basic step in the process of forecasting. The evolutionary model has two types, In my research work used two forms of machine learning based algorithms that are-

Decision Tree- The Decision Tree can be a supervised learning method that will be used to differentiate between Regression problems, but is especially preferred to solve Separation problems. Decision Tree can be a tree-shaped divider, where the inner nodes represent the elements of the database, the branches representing the selected rules and all the leaf nodes represent the result.

Artificial Neural Network (ANN) - ANN design concept to mimic the way the human brain works. ANN consists many layer that is input layer, several hidden layers, and an output layer. Units with adjacent layers are fully connected. The ANN contains a large number of units and can measure the activities of the opposition; hence it has the right balance, especially in offline activities. Due to the complex model structure, ANN training is time consuming. It is noteworthy that ANN models are trained by a built-in algorithm that will not be used to train deep networks.

Evaluation criteria-The goal of this study is to estimate the prediction of thyroid disease by utilization of various machine learning approaches. The outcomes from different models and their ensemble models are analyzed and compared on a many evaluation criteria, such as

Accuracy - it's the foremost intuitive performance measure. it's simply a ratio of correctly predicted observation to the overall observations.

$$\text{Accuracy} = TP+TN/TP+FP+FN+TN$$

Precision - it's the ratio of correctly predicted positive observations to the entire predicted positive observations.

$$\text{Precision} = TP/TP+FP$$

Recall - it's the ratio of correctly prediction of positive observations to the all observations in actual class - yes.

$$\textbf{Recall} = TP/TP+FN$$

IV. RESULT: Thyroid Data set contains patient health features like age, sex, on thyroxine, quer on thyroxine, on anti-thyroid medication, thyroid surgery, query hypothyroid, query hyperthyroid, pregnant, sick, tumor, lithium, goitre, TSH measured, TSH, T3 measured, T3, T4 measured,T4,TT4 measured,TT4, T4U measured, T4U, FT1measured, FT1, TBG measured, TBG. 1030 patient records were taken for performance with 9 selected features. Following Table describes the accuracy of both classification technique for Thyroid dataset for selected features. Artificial Neural Network classifier perform best result compare than decision tree classifiers i.e, 97% accuracy

Classifier	**Accuracy**
ANN	**97.1%**
Decision Tree	**93.7%**

Table 2: Accuracy table

V. CONCLUSION

In this paper, two classification techniques are study for the prediction of diseases. The classification model is generated based on the training data, and the data is tested though predictions in the prediction stage. This research is focus on accurate prediction model will be built by choosing an effective classification technique. comparison between classification algorithms based on the accuracy and confusion matrix than Effective classifier is identified by the performances. In this paper, two classification techniques were implemented for the prediction of diseases. The classification model is generated based on the training data, and the data is tested though predictions in the prediction stage. Thyroid, dataset were tested using classification algorithm using Python environment. It is clear that ANN classification Algorithm has given the best accuracy results when compared to Decision tree.

VI. REFERENCES

1. Ambika Gopalakrishnan Unnikrishnan and Usha V. Menon, "Thyroid disorders in India: An epidemiological perspective", Indian Journal of Endocrinology and metabolism, 2011 Jul; 15(Suppl2): S78–S81.
2. Anupam Shukla , Prabhdeep Kaur ,"Diagnosis of thyroid disorders using ANN" International advance computing, conference , IEEE 2009.
3. Fatemeh Saiti,Afsaneh Alavi and Naini Mahdi Aliyari, Shoorehdeli," Thyroid Disease Diagnosis Based on Genetic Algorithms Using PNN and SVM", 3rd international conference on bioinformatics and biomedical engineering, 2009.
4. G. Rasitha Banu ," Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique ",Communications on Applied Electronics (CAE) – ISSN : 2394- 4714 Foundation of Computer Science FCS, New York, USA Volume 4– No12, January 2016.
5. Gurmeet Kaur and Er. Brahmaleen Kaur Sidhu, "Proposing 4] Gurmeet Kaur and Er. Brahmaleen Kaur Sidhu, "Proposing Efficient Neural Network Training Model for Thyroid Disease Diagnosis.", International Journal For Technological Research In Engineering Volume 1, Issue 11, ISSN (Online): 2347 - 4718, pp. 1383-1386, July-2014.
6. https://github.com/renatopp/arff datasets/blob/master/classification/hypothyroid.arff .
7. Jamil Ahmed and M.Abdul Rehman Soomrani ,"TDTD: Thyroid disease type diagnosis"2016 international conference on intelligent systems engineering(ICISE),pp.44- 50, IEEE, 2016,
8. K. Rajam, R. Jemina Priyadarsini, "A Survey on Diagnosis of Thyroid Disease Using Data Mining Techniques", International Journal of Computer Science and Mobile Computing, IJCSMC, Vol. 5, Issue. 5, May 2016, pg.354 – 358.
9. Liyong Ma,Chengkuan Ma,Yuejun Liu,Xuguang Wang,"Thyroid Diagnosis from SPECT Images Using Convolutional Neural Network with Optimization"Computational Intelligence andNeuroscience,Volume2019,https://doi.org/10.1155/2019/6212759.
10. Marissa Lourdes De Ataide1, Amita Dessai2Thyroid Disease Detection using Soft Computing Techniques, , International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072.
11. 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.
12. S. Bagcchi, "Hypothyroidism in India: More to be done," The Lancet Diabetes & Endocrinology, vol. 2, no. 10, p. 778, 2014.
13. Sayyad Rasheeduddin, Kurra Rajasekhar Rao,,"Extreme Learning Machine for Thyroid Nodule Classification with Graph Cluster Ant Colony Optimization Based Feature Selection", International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-2, July 2019.