Data Mining Classification on Hypo Thyroids Detection: Association Women Outnumber Men

Suwarna Gothane

Abstract: ***Thyroid diseases are common worldwide and affecting health life. Health care is an inevitable task to be done in human life. In this paper, we have made an attempt to diagnosis hypothyroid. This paper analyzes few essential parameters affecting thyroid. Data mining acts as a solution to many thyroid healthcare problems. To overcome the problem we have come here with a novel solution approach to identify key factors affecting hypothyroid using WEKA tool 3.8. This paper presents thyroid data analysis, classification and prediction. The result of the proposed work used for the forthcoming identification to keep track on important factors affecting hypothyroid.***

Keywords: Data Mining, Hypo Thyroidism, Prediction

I. INTRODUCTION

Thyroid one of the most popular disease across the world is not exception for the India. Various studies indicated 42 million people in India undergoes through thyroid disease. Ambika Gopalakrishnan et.al, identified various thyroid related disorder in India includes hypothyroidism, hyperthyroid, goiter, iodine shortage diseases, Hashimoto's thyroiditis, thyroid cancer. Author worked on determining thyroid hormones normal indication range, particularly during pregnancy on Indian dataset[1].Report identified prevalence of Hypo Thyroidism is around 11% in India.[13]

In this paper, the work is furnished such as in the section – II the outcomes from the parallel researches are discussed. In the section – III Proposed Research Work Approach is discussed. In the next section, Section – IV implementation is discussed, Followed by the results are furnished and explained in the Section – V and the work presents the final conclusion in the Section – VI.

II. RELATED WORK

Sayyad Rasheeduddin et.al used unsupervised Graph Clustering ant Colony Optimization based Extreme Machine Learning technique to detect threat of thyroid. This approach identified factors affecting risk of thyroid. Author found approach superior than univariate logic. The ultrasound technique used for identification found popular, affordable and efficient [2]. Liyong Ma, Chengkuan Ma et.al proposed proficient way using convolutional neural network to detect disease illnesses on SPECT datasets. Suggested approach result worked better than existing methods [3].

K. Rajam et.al noticed 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 [4].Marissa Lourdes De Ataide et.al applied a two multilayer perceptron classifier for classifying thyroid diseases into three classes as euthyroid, 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 [5].To reduce problem [6],[7],[8],[9], applied neural network for analyzing thyroid problem. Fatemeh Saiti et.al, applied Genetic Algorithms Using Support Vector Machine for thyroid verdict [10]. G. Rasitha Banu predicted problem using Linear Discriminant Analysis (LDA)- Data Mining approach[11].

Revised Manuscript Received on February 01, 2020.

Dr. Suwarna Gothane, Professor, Department of Computer Science and Engineering, MLR Institute of Technology, Hyderabad, India.

III. PROPOSED WORK

Various author applied supervised and unsupervised approaches such as machine learning, neural network, Naïve bayes, decision tree, back propagation, multilayer perceptron, Genetic Algorithm and Support vector machine. Further future work scope is available to expand on larger size datasets. Inadequate data blocks need detail classification and diagnosis of thyroid diseases. The data obtained from diagnoses by professional physicians can serve as fundamental correct source of data. We need sophisticated system to identify thyroid and also to recommend level that can further leads to serious problems like cancer. We are using in our proposed work data mining approach. We have taken datasets hypothyroid from website source github. Thus the proposed novel developed architecture for Hypo Thyroid Detection is available in Fig. 1.

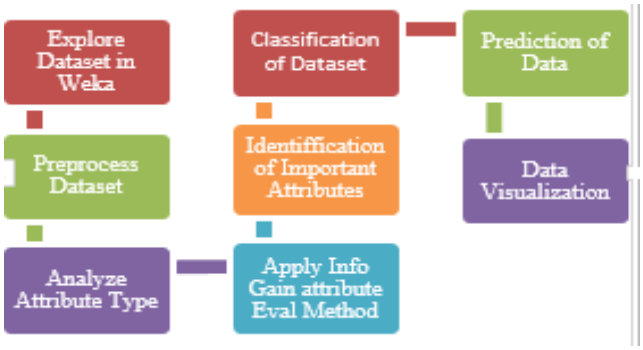

Fig .1.Proposed Architecture for Hypo Thyroid Detection and classification.

Our proposed system architecture Hypo Thyroid Detection and classification consist of following

A. Explore Dataset in WEKA Tool

We used free Hypothyroid dataset from github library and explore it in WEKA Data Mining Tool.

B. Preprocess dataset

Here dataset is preprocesses for all attributes. Preprocessing is the basic step in data mining whether we perform

check for attributes and instances.

C. Analyze Attribute Type

We perform check for all attribute type as nominal numeric or class type.

D. Apply Info Gain attribute Eval Method

We applied Info Gain Attribute Eval Method and identified which attribute are more dangerous for Hypothyroid

E. Identify Important Attributes

Ranking of Attribute is considered here using Ranker Algorithm to identify important attribute.

F. Classification of Dataset

We have classified dataset under 4 labels such as negative, compensated_hypothyroid,primary_hypothyroid,secondary_ hypothyroid to understand severity of disease and further to take suitable measure and care

G. Prediction of Dataset

The obtained parameter can further use to predict Hypothyroid for the future to keep track on important factors affecting hypothyroid.

H. Data Visualization

From the retrived result BarGraph, PieChart used as a medium Visualization

IV. IMPLEMENTATION OF PROPOSED WORK

We considered dataset of 30 attributes which includes age, sex, on thyroxine, query on thyroxine ,on antithyroid medication, sick, Pregnant, thyroid surgery,I131 treatment, query hypothyroid, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH measured,TSH,T3 measured,T3,TT4 measured,TT4,T4U measured,T4U,FTI measured, FTI,TBG measured, TBG, referral source, Class. We focused on Identification of Attribute Type, Identification of Important Attribute contributing to thyroid, Classify the dataset using ZeroR Classifier.

Identification of Attribute Type

The first outcome of this work is to identify attribute under 3 category nominal, numeric and class.

Nominal Attributes of the dataset is shown in Table-I:

Table-I: Nominal Attribute Identification

Sr. No	Attribute Name	Attribute Category	Number	Label	Count
1	Sex	Nominal	1	F	2480
			2	M	1142
2	on thyroxine	Nominal	1	f	3308
			2	t	464
3	query on thyroxine	Nominal	1	f	3722
			2	t	50
4	on antithyroid medication	Nominal	1	f	3729
			2	t	43
5	Sick	Nominal	1	f	3625
			2	t	147
6	Pregnant	Nominal	1	f	3719
			2	t	53
7	thyroid surgery	Nominal	1	f	3719
			2	t	53
8	I131 treatment	Nominal	1	f	3713
			2	t	59
9	query hypothyroid	Nominal	1	f	3538
			2	t	234
10	query hyperthyroid	Nominal	1	f	3535
			2	t	237
11	Lithium	Nominal	1	f	3754
			2	t	18
12	Goiter	Nominal	1	f	3738
			2	t	34
13	Tumor	Nominal	1	f	3676
			2	t	96
14	Hypopituitary	Nominal	1	f	3771
			2	t	1
15	Psych	Nominal	1	f	3588
			2	t	184
16	TSH measured	Nominal	1	t	3403
			2	f	369
17	T3 measured	Nominal	1	t	3003
			2	f	769
18	TT4 measured	Nominal	1	t	3541
			2	f	231
19	T4U measured	Nominal	1	t	3385
			2	f	387
20	FTI measured	Nominal	1	t	3387
			2	f	385
21	TBG measured	Nominal	1	f	3772
			2	t	0

Numeric Attributes of the dataset is shown in Table-II:

Table-II: Numeric Attribute Identification

Sr.No	Attribute Name	Attribute Category	Statastic	Value
1	Age	Numeric	Minimum	1
			Maximum	455
			Mean	51.736
			StdDev	20.085
2	TSH	Numeric	Minimum	0.005
			Maximum	530
			Mean	5.087
			StdDev	24.521
3	T3	Numeric	Minimum	0.05
			Maximum	10.6
			Mean	2.013
			StdDev	0.827
4	TT4	Numeric	Minimum	2
			Maximum	430
			Mean	108.319
			StdDev	35.604
5	T4U	Numeric	Minimum	0.25
			Maximum	2.32
			Mean	0.995
			StdDev	0.195
6	FTI	Numeric	Minimum	2
			Maximum	395
			Mean	110.47
			StdDev	33.09
7	TBG	Numeric	Minimum	NaN
			Maximum	NaN
			Mean	NaN
			StdDev	NaN
8	referral source	Numeric	SVHC	386
			other	2201
			SVI	1034
			STMW	112
			SVHD	39

Classification attribute of the dataset is shown in Table-III

Table III: Classification Attribute Identification

Sr. No	Classification	Attribute Type	Classification Type	Classification Result
1	Class	Numeric	negative	3481
			compensated_ hypothyroid	194
			primary_ hypothyroid	95
			secondary_ hypothyroid	2

A. Identification of Important Attribute

The second outcome of this work is to identify Important attribute. We applied InfoGain Attribute Eval function on dataset which consist of 3772 instances and 30 attributes.Considering all train data with Information Gain Ranking Filter we got result of Selected attributes: 18,26,22,20,17,3,29,10,24,21,2,16,19,6,7,8,13,23,25,5,4,11, 12,14,9,15,28,27,1 : 29 Ranked attributes shown in Table IV:

Table-IV: Ranked Attribute using InfoGain Attribute Eval Method

0	1	age	29
0.0020247	2	sex	11
0.0103441	3	On thyroxine	6
0.0005113	4	Query onthyroxine	21
0.0006123	5	On antithyroidmedication	20
0.0017492	6	sick	14
0.0016395	7	pregnant	15
0.0011337	8	Thyroid surgery	16
0.000044	9	I131 treatment	25
0.0045771	10	Query hypothyroid	8
0.0004252	11	Query hyperthyroid	22
0.0001803	12	Lithium	23
0.001049	13	Goiter	17
0.0000776	14	Tumor	24
0.0000307	15	hypopituitary	26
0.0019482	16	Psych	12
0.0119506	17	TSH measured	5
0.3318159	18	TSH	1
0.0018619	19	T3 measured	13
0.0398198	20	T3	4
0.0030541	21	TT4measured	10
0.1322953	22	TT4	3
0.0006716	23	T4U measured	18
0.0040942	24	T4U	9
0.142952	26	FTI	19
0	27	TBG measured	2
0	28	TBG	28
0.0060967	29	Referral source	27

C. Classify the dataset

The third outcome of this work is to perform classification of dataset. We used ZeroR classifier and retrieved results on dataset name hypothyroid with 30 attributes and 3772 data values. Results are observed in 0.02 seconds. We found properly classified values as 3481 and improperly classified instances as 291.We obtained 92.2853 % of accuracy positive and 7.7147% negative accuracy of classification.

V.RESULT ANALYSIS

A. InfoGain Attribute Evaluation Results

The comparison is presented in Graphical Format

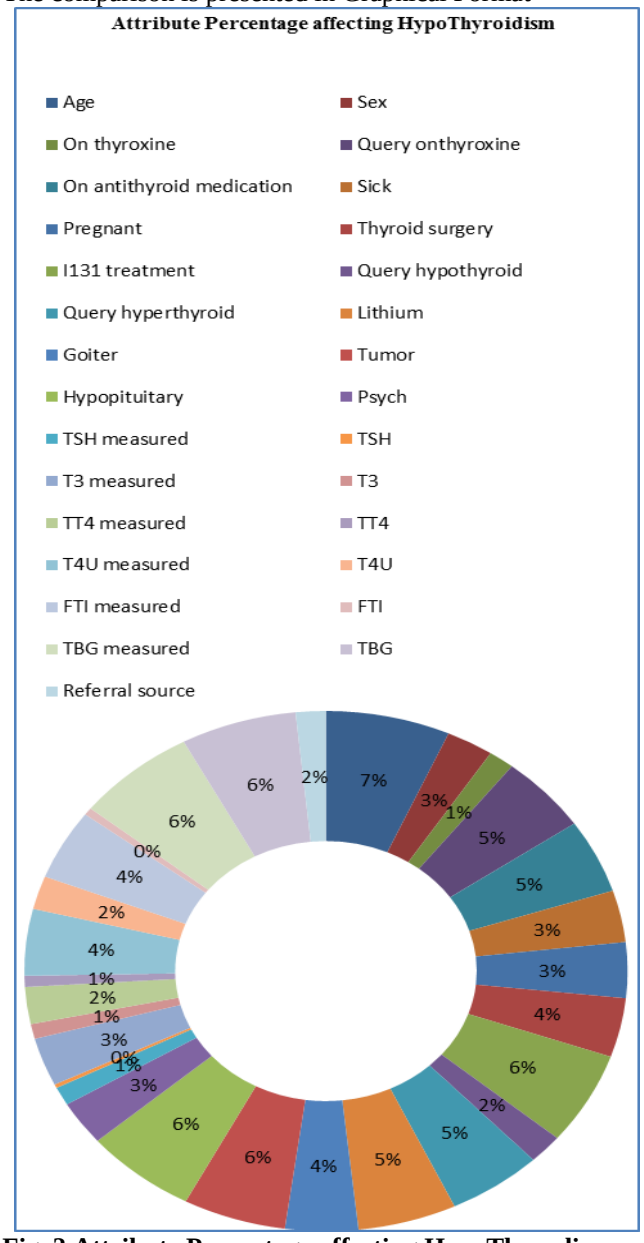

Fig .2.Attribute Percentage affecting HypoThyrodism

The results are been visualised graphically to understand the classification Fig. 3.

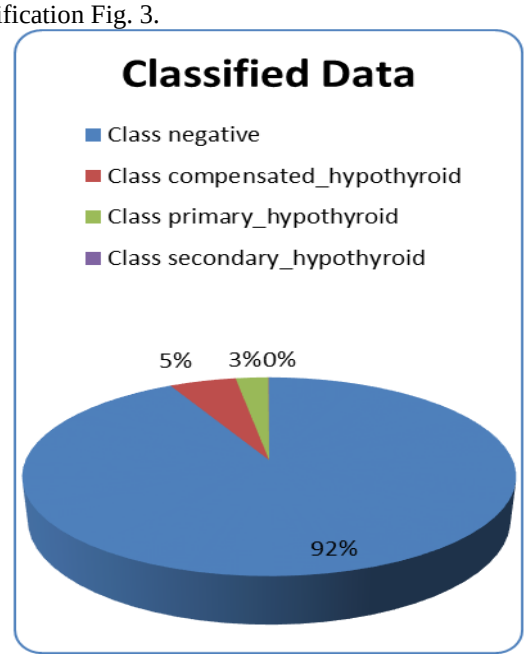

Fig.3.Classification of Data

C. Female and Mail Classification and graph

The graph shows female are more prone to hypothyroidism than male depicted in Fig.4.

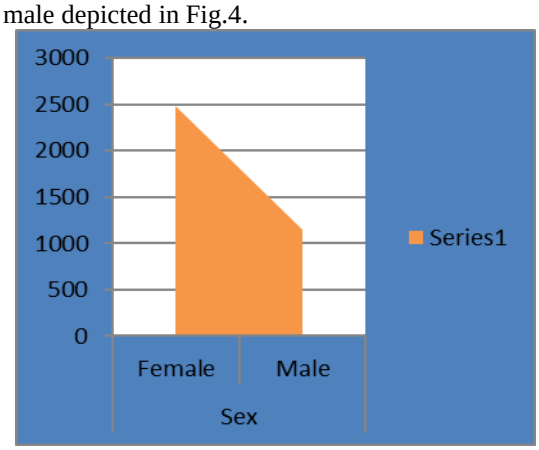

Fig. 4. Female and Male Classification

VI. CONCLUSION

We used WEKA tool 3.8 and perform classification using Zero R on dataset Correctly Classified Instances 3481 i.e 92.2853 % and Incorrectly Classified Instances 291 7.7147 %.Under 4 category classified data with label negative3481compensated_hypothyroid194,primary_hypothyroid95secondary_hypothyroid 2.This paper presents thyroid data analysis, classification and prediction. The outcome of the extracted data can be analyzed for the future to keep track on important factors affecting hypothyroid. Here we have also identified that Female contributes more comparatively than male towards hypothyroid. The obtain information serves useful and plays vital role in future for women to be careful towards thyroid.

ACKNOWLEDGMENT

Apart from the efforts of me, the success of paper depends largely on encouragement and guidelines of many others. I take this opportunity to express my profound gratitude to MLR Institute of Technology College Management for motivating me and for providing me all the facilities required for this work. I am deeply indebted to Chairman Shri. M.Laxman Reddy, Secretary Shri. Marri Rajashekar Reddy, Director Dr. K. Srinivas Rao, Dean Dr. Purna Chandra Rao, MLRIT who always has been a constant source of inspiration for me.

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. 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.
3. 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.
4. 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.
5. 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.
6. 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, DOI:10.1109/INTELSE.2016.7475160
7. Anupam Shukla , Prabhdeep Kaur ,"Diagnosis of thyroid disorders using ANN" International advance computing, conference , IEEE 2009.
8. 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.
9. 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.
10. 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.
11. 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.
12. https://github.com/renatopp/arff datasets/blob/master/classification/hypothyroid.arff
13.]S. Bagcchi, "Hypothyroidism in India: More to be done," The Lancet Diabetes & Endocrinology, vol. 2, no. 10, p. 778, 2014.

AUTHORS PROFILE

Dr. Suwarna Gothane presently working as Professor in MLR Institute of Technology, Hyderabad, Telangana, INDIA. She received her Ph.D (CSE) from Sant Gadge Baba Amravati University, Amravati in year 2019, M.E. (CSE) degree from P.R.M.I.T&R, Amravati in the year 2012 and B.E. (CSE) degree from H.V.P.M C.O.E & Technology, in the year 2006. Her area of interests are Data Mining, Machine learning.