Supervised Machine Learning Models For Classification Of Thyroid Data

D.Hemalatha , S. Poorani

Abstract: Indecorous activation of thyroid glands becomes major issue of concern among Indian women. Hyperthyroidism and hypothyroidism are the two major thyroid disorders that should be treated early. Hyperthyroidism occurs due to over secretion of hormones than the need of the body. Hypothyroidism causes due to surplus exertion of hormones from the thyroid gland. T3, T4 and TSH hormones play a significant role in functioning of the thyroid gland. Various studies have been done to predict the thyroid disorder. The key objective of this research work is to predict the type of thyroid disorder using supervised ML techniques.

Index Terms: Classification, Decision Tree, KNN, Machine Learning, Naive Bayes, SVM,

1. INTRODUCTION

Women are generally affected by thyroid disorder, which lay the foundation for various kinds of health problems such as hormonal imbalances, weight gain, weight loss and others. Men are also at risk, but the chances of suffering from thyroid diseases are considered to be few as related with women. Common statistics reports Women may have a chance of thyroid up to eight times higher than men. Globally one third of people who have thyroid disorder are Indian. A thyroid disorder affects 42 million people in India. Northern part of India noted the maximum cases of hypothyroidism and hyperthyroidism is seen in the south and the west zones [1]. Hypothyroidism is a very serious problem in India, where 1 in every 10 men & women suffer from hypothyroidism [2]. To diagnose thyroid diseases, doctors use a medical history, physical exam, and thyroid tests. They sometimes also use a biopsy. Treatment depends on the problem, but may include medicines, radioiodine therapy, or thyroid surgery. The occurrence of hypothyroidism in India is 11%, compared with only 2% in the UK and 4·6% in the USA. This is possibly linked to long-standing iodine deficiency in the country, which has only been partly corrected over the past 20 years. Apart from iodine deficiency, environmental factors can play a part in hypothyroidism. Fuzzy rule-based system has been used for thyroid prediction [3]. Numerous classification methods like K-nearest neighbour, Naive Bayes and SVM are used for prediction of thyroid disorder. The K-nearest neighbour gives better accuracy than Naive Bayes [4]. Various data mining techniques like C4.5, KNN, LDA, Neural network and random forest methods are used to classify the hypothyroid datasets for the early prediction of thyroid disorder. Few of these algorithms are combined with kfold cross validation, which gives the exact accuracy [5]. Optimal feature selection and kernel-based classifier process has used to classify the thyroid data. Classification process is combined with gray wolf optimization to enhance the performance [6]. The focal principle of this study is to perceive, and examine thyroid Disease. In this study, the four classification models Naïve bayes, Decision tree, KNN and SVM are trained with same thyroid dataset. Then all models are tested with new test data and finally, a proportional study is performed to get appropriate classifier.

- *Ms.D.Hemalatha is currently working as Assistant Professor in the Department of Computer Technology, Kongu Engineering College, affiliated to Anna University Tamilnadu, India. E-mail: hema.arun2011@gmail.com*
- *Ms. S. Poorani is currently working as Assistant Professor in the Department of Computer Technology, Kongu Engineering College, affiliated to Anna University Tamilnadu, India.E-mail : vspoorani@gmail.com.*

2 RELATED WORKS

Negar Asaad Sajadi et al [3] proposed an expert system for prediction of hypothyroidism. They used fuzzy rule based system and gives 97% of accuracy for thyroid disorder prediction. They used real dataset collected from Shahid Beheshti Hospital in Hamadan west of Iran. The dataset contains 305 instances, which contains three classes like normal, subclinical hypothyroidism and hypothyroidism. Attribute includes demographic, symptoms and laboratory tests. According to their system, fuzzy rule based classification gives greater precision rate compared to earlier findings of thyroid disorder prediction. For the subclinical hypothyroidism, fuzzy classifier gives improved performance than the logistic regression model. Khushboo Chandel et al [4] used classification techniques like K-nearest neighbour, Naive Bayes and SVM. They discussed about data mining techniques used for disease detection used over the last 15 years for detecting several diseases. The accuracy of KNN is better than Naive Bayes to detect thyroid disorder. The parameters used to classify thyroid disorder are TSH, T4U and goiter. They used dataset from Knowledge Extraction Evolutionary Learning (KEEL) Repository and considered 7200 instances for analysis of thyroid prediction. The KNN produce accuracy of 93.44% where Naïve Bayes gives 22.56% accuracy which is relatively lesser than KNN. Cross-validation was applied to achieve the best results in order to estimate the statistical performance. They suggest other classifiers like ACP, evolutionary approach and swarm intelligence algorithm for further work. Roshan Banu D and Sharmeli K C [5] used Random forest algorithm and SVM to predict the thyroid disorder. They used TBG as predicting value and also the attributes like TSH, T3 and T4U. Random Forest algorithm gives more accuracy to predict thyroid disorder compared to SVM. RF gives accuracy of 70% by taking the TBG_MEASURED attribute. K. Shankar et al [6] use optimal feature selection kernel-based classifier process for the

prediction of data thyroid disorder. Dataset is collected from UCI repository. The proposed work consists of three stages: pre-processing, feature selection and classification. Classification is done using multi kernel SVM. To enhance the performance of classifying process feature selection is done with improved gray wolf optimization. They took 29 attributes and the feature selection is carried out with improved gray wolf

optimization. The proposed classifier model uses Multi Kernal SVM to classify the data as hypothyroidism , hyperthyroidism or normal. The proposed system for thyroid classification gives accuracy of 97.49, sensitivity of 99.05 and specificity of 94.5%. Ammulu K and Venugopal [7] T predicted hypothyroidism disorder using random forest approach. The dataset used in this work is collected from UCI. The concert measure is evaluated using confusion matrix. The Weka tool was used for experiments. K.Geetha, S. Santhosh baboo [8] classified thyroid disease using differential evolution with support vector machine. The database is taken from UCI repository with consists of 21 attributes and 7200 instances. The attributes consists of categorical data and real data. Using Hybrid Differential Evolution Kernel Based Naive Based algorithm these attributes are then optimized to 10 attributes. The dataset is pre-processed .The pre-processed data is fed into a hybrid algorithm called differential evolution. The fitness is checked by providing data to bayesian classifier with Kernel function and error-stabilization is calculated to deliberate the fitness. After stabilization is achieved , the data is classified into 3 classes as Hypo Thyroid, Hyper Thyroid and Normal. The optimized dataset is given to SVM Classifier where RBF is used to predict the thyroid disorder, which produces accuracy of 99.89%.In the existing works, three different datasets like real dataset which was collected from hospital [3], KEEL repository [4] and UCI-data [6, 7, 8] were used with fuzzy-classification, KNN & NB and SVM, RF & kernel-based NB respectively. The existing SVM with UCI used 29 attributes and KNN & NB used KEEL data. In the proposed system UCI data with 21 attributes are analyzed with Decision tree, KNN, NB and SVM.

3 METHODOLOGY

The dataset used in this work is taken from UCI-library[10] which has 21 attributes and 7200 instances. It includes three labels 1-represents normal, 2-represents hyperthyroidism &3-represents hypothyroidism. The supervised ML methods rely on labelled input data to gain knowledge about a function which generates proper output while new-fangled data is given without label. Naive Bayes is a keen erudition classifier and certainly, it is hasty and well-known method for multi-class classification. So that in real-time, the predictions can be made with this algorithm. Here the likelihood of multiple-classes of the objective trait is predicted. SVM classifies the data based on hyperplanes and they are innately two-class classifiers. A superior substitute is endowed with the erection of multiclass SVMs, in which we construct a two-class classifier. Decision tree erudition is a dominant classification model. To construct a superior generalization, the tree strives to deduce a rip of the data used to train the model based on the principles of the obtainable traits. Naturally, this algorithm can solve two-class and multiclass problems. The KNN algorithm presumes that akin things subsist in close propinquity. That is, analogous objects are close to each other. KNN confines the inspiration of resemblance with mathematics like manipulating the space connecting two ends in a graph. The k value used here is 13.

4 RESULTS AND DISCUSSION

In this investigation, totally four classifiers are attempted to perform multiclass classification. All the four classifiers uses the same dataset[10]. The models are trained with 75% of whole data and the lingering 25% are used to estimate the concert of classifiers. The data has three classes in total, in which label 3 is highly distributed than 1 and 2. The confusion matrix and accuracy of all the models are given in Table I and

TABLE 1
CONFUSION MATRIX

Classifier	Actual	Prediction		
		1	2	3
Naive Bayes	1	42	0	0
	2	16	76	0
	3	185	1433	48
SVM	1	27	4	11
	2	1	17	74
	3	1	0	1665
K-nearest Neighbour	1	21	0	0
	2	0	34	0
	3	0	2	663
Decision tree	1	30	1	3
	2	12	86	9
	3	0	5	1654

Table II. Among four classifiers accuracy of naïve bayes classifier is very low. The other classifiers give better accuracy. The KNN performed best than others. Eventhough Decision tree and SVM provides good accuracy, KNN will be the appropriate model for multiclass classification of thyroid data.

5 CONCLUSIONS

To predict thyroid disorder, the ML techniques NB, DT, KNN and SVM are examined for three-class classification and all classifiers produce good results except NB. In future these algorithms can be implemented for prediction of thyroid disease with more real data related with thyroid and with binary as well as multiple classes.

TABLE 2
PERFORMANCE OF ALL CLASSIFIERS

CLASSIFIER	ACCURACY(%)
NAIVE BAYES	9.2
SVM	94.94
DECISION TREE	98.33
KNN	99.72

REFERENCES

[1]https://economictimes.indiatimes.com/magazines/panache/over-30-indians-suffering-from-thyroid-disorder-survey/articleshow/58840602.cms May 25, 2017.
[2] https://www.downtoearth.org.in/news/health/1-in-10-indians-have-hypothyroidism-61693,September 2018.
[3] Negar Asaad Sajadi, Shiva Borzouei, Hossein Mahjub and Maryam Farhadian, Diagnosis of hypothyroidism using a

fuzzy rule-based expert system, Clinical Epidemiology and Global Health, 2018, doi: https://doi.org/10.1016/j.cegh.2018.11.007.

[4] Khushboo Chandel, Veenita Kunwar, Sai Sabitha, Tanupriya Choudhury & Saurabh Mukherjee, A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques, CSIT, December 2016 ,pp.313–319.

[5]Roshan Banu D , Sharmeli K C , Classification Model Using Random Forest and SVM to Predict Thyroid Disease, International Journal on Future Research in Computer Science & Communication Engineering, ISSN: 2454-4248, Volume: 4 Issue: 2,2018, pp.85 – 87.

[6] K. Shankar et al, Optimal feature-based multi-kernel SVM approach for thyroid disease classification, The Journal of Supercomputing, 2018, https://doi.org/10.1007/s11227-018-2469-4

[7] Ammulu K., Venugopal T, International Journal for Innovative Research in Science & Technology, 2017, Volume 4 , Issue 2,pp. 208-212.

[8] K.Geetha, capt. S. Santhosh baboo, Efficient thyroid disease classification using differential evolution with svm, Journal of Theoretical and Applied Information Technology, 30th June 2016. Vol.88. No.3,pp.410-421.

[9] Maurya H, Thyroid Function Disorders among the Indian Population, Annals Thyroid Res, Austin Publishing Group , 2018; 4(3): 172-173.

[10]https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease