International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

[1]Thyroid Prediction Using Machine Learning Algorithms

Sumit Mhaiskar
Bharati Vidyapeeth's Institute of Management and Information Technology,
Sector-8, CBD Belapur, Navi Mumbai, Maharashtra-400614.

Abstract—The paper has been written with the aim of making a machine learning model that will predict thyroid disease with better accuracy. Among different machine learning algorithms, we are using supervised machine learning algorithms like Random Forest, Support Vector Machine (SVM), and Logistic Regression to predict thyroid disease and to evaluate their performance in terms of accuracy. For prediction of the thyroid disease, we are using the dataset from the UCI repository. Data from the UCI repository has been combined and cleaned to detect whether the person is suffering from thyroid or not.

Index Terms—Thyroid Disease, Machine Learning, SVM, Random Forest, Logistic regression

I. INTRODUCTION

Thyroid disease is a commonly known disease that affects human health and a person at least out of 10 suffer from this disease. The statistics say that thyroid in India is rising and many people suffer from hypothyroidism. It is a condition in which the thyroid gland does not produce enough thyroid hormones to meet the needs of the body. Thyroid disease can be commonly seen in women between the age of 18-35 which is the most crucial period of their life.

The thyroid gland is a small organ that is located in the neck. Its shapes are like a butterfly, which is small in the middle and has two wide wings that extend at the side of the throat.[8] We have many glands in our body where that are responsible for creating and releasing the substances that help the body to perform many vital functions.

When the thyroid gland doesn't work properly, it can affect our entire body. If the body produces too much thyroid hormone, then the body develops a condition called hyperthyroidism and if the body produces very less hormone, then the body develops a condition called hypothyroidism.[10] There are two hormones that are produced by the thyroid glands that is T3 (Triiodothyronine) and T4 (Thyroxine).

T3 and T4 together can affect almost every cell in the body. The thyroid gland produces more T4 hormone than T3 but T4 gets converted to T3 when the hormone reaches to the cells and tissues that are present inside the body.[13] Therefore, T4 hormone is the most important hormone to measure when testing for the thyroid problems.[6] The T3 hormone contains three iodine molecules in its structure, and it is more metabolically active when it controls the body metabolism, temperature, and the digestive system. Iodine is considered as the main building block of the thyroid gland. The thyroid gland is under the control of the pituitary gland. When the level of T3 and T4 hormone drops too low, the pituitary gland produces Thyroid Stimulating hormone that is also known as TSH that stimulates the thyroid glands to produce more hormones. As the data is increasing day by day and the technology is improving more, research in the healthcare domain is also increasing. It might be difficult to handle a lot of amounts of patient's data and that's where machine learning comes into the picture. Machine learning is a way that helps in early prediction and leads to proper diagnosis of disease.

Various machine learning algorithms helps to discover the hidden patterns, train and building of a model and make predictions by learning from training data. Various supervised machine learning algorithms helps to train the machine using well labelled training data and basis on the data, the machine predicts the output.[9]

II. LITERATURE STUDY

There has been a lot of work in the field of healthcare domain using multiple machine learning algorithms. Many people have used many different types of data mining techniques to predict the disease. The predictive analysis of machine learning can also help users to get the personalized treatment. Machine learning can assist doctors by helping them for diagnosing the disease and thereby reducing their burden.

"Thyroid Detection Using Machine learning" Published Online January 2021 in IJEAST, in that they have used various machine learning algorithms like SVM (Support Vector Machine), decision tree, logistic regression KNN (K-Nearest Neighbours), ANN (Artificial Neural Network) to predict the chances of a person having a thyroid disease. They have also

created a web application to get the data from the users to predict the type of the disease.

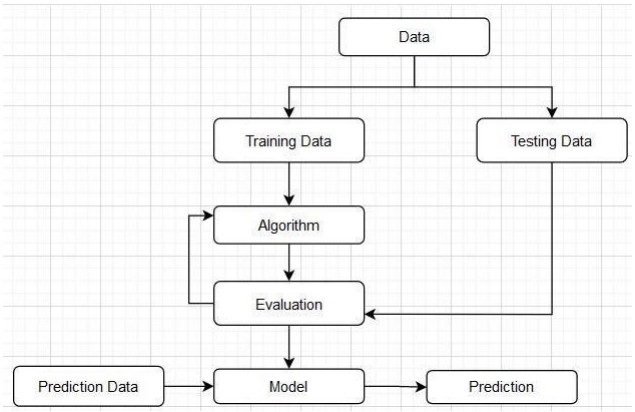
They have taken the dataset from the UCI repository where the dataset has two labels that is hyper and hypo. After training and building the model, they have achieved an accuracy score of 93.84% using KNN algorithm, 95.38% using the SVM algorithm, 75.38% using the ANN algorithm, 92.3% using the decision tree algorithm, 96.92% using the logistic regression algorithm. Since they have achieved the highest score using the logistic regression algorithm, they have considered that as their prediction model.

" Interactive Thyroid Disease Prediction System Using Machine Learning Technique" in their work they have use the machine learning algorithms like SVM (Support Vector Machine), KNN (K-Nearest Neighbours), Decision Trees for predicting the thyroid disease using the dataset from the UCI machine learning repository. They have taken the dataset from the UCI repository, and they have achieved an accuracy score of 98.62% using KNN algorithm, 99.63% using the SVM algorithm, 97.50% using the ANN algorithm, 75.76% using the decision tree algorithm. They have achieved the highest score using the SVM algorithm.

" Thyroid Disease Prediction Using Machine Learning Approaches" published in the National Academy of Science, India 2020 they have use the dataset from UC Irvin knowledge discovery in databases archive. They have used the algorithm like KNN (K-Nearest Neighbours), decision tree and logistic regression. They have used all these algorithms to apply classification on the data.

They have achieved an accuracy score of 96.875% using KNN algorithm, 87.5% using the decision tree algorithm, 81.25% using the logistic regression classifier. They have achieved the highest score using the KNN algorithm. The main objective of this study is to develop a system which can predict if the person is suffering from a thyroid or not. Along with to predict the disease with a smaller number of parameters. Also, to provide an efficient solution for this healthcare problem. We will be using multiple machine learning algorithms to predict the disease like KNN, Random Forest etc.

III. RESEARCH METHODOLOGY



For predicting the thyroid disease dataset of thyroid is required for analysing and predicting the disease. We will analyse the dataset using various supervised machine learning algorithms. Based on the accuracy of different models, the algorithm which will give the highest accuracy score will be chosen to fetch the result.

The dataset is taken from the UCI repository. The dataset needs to be checked for null or empty values or even for the unnecessary values. Then this type of values gets removed from the data and thereby the data gets clean. During the data cleaning process, the parameter which are needed for the prediction of the disease only those parameters are kept, and rest of the parameters are dropped.

The data which is cleaned is then used as training and testing data, which is given as input to the algorithms. The algorithms extract the features from datasets for classification of the data according to the labels. Then to check if the prediction is accurate or not the test data is fed to the algorithm along withit accuracy score for every model is also calculated to understand the accuracy of different models.

A. Attributes used for diagnosis of the thyroid disease

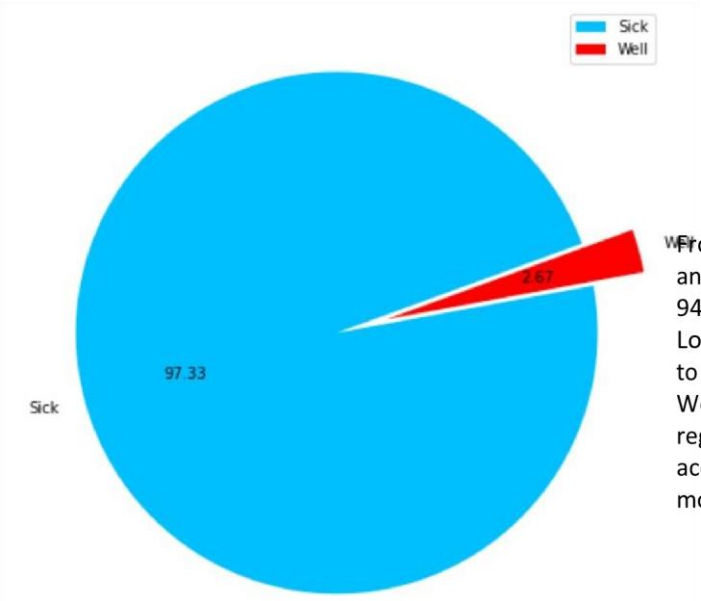
Attributes	Description
Age	In years
Sex	Male or female
TSH	Thyroid-Stimulating Hormone
T3	Triiodothyronine
TBG	Thyroid binding globulin
T4U	Thyroxin utilization rate
TT4	Total Thyroxin
FTI	Free Thyroxin Index

Fig. 1. Table 1: Attributes for feature selection

The attributes that are important for the prediction of the diagnosis are listed below. Almost most of the research have been done using the following attributes.

B. Performance Study of the proposed algorithm

Random Forest: It is a machine learning algorithm that can be used to solve both classification and regression problems. It is based on the ensemble learning that helps in combining multiple classifiers to solve a complex problem to improve the performance of the model. It takes number of decision tree on the subsets of the dataset and takes the average to improve



the predictive accuracy of the dataset. The algorithm takes the prediction from each decision tree and based on the majority of the votes the algorithm predicts the final output. Logistic Regression: This algorithm is used for predicting the categorical dependent variable using different independent variable. This algorithm is used to solve the classification problem. In this algorithm we fit the "S" shaped logistic function that will predict the values like 0 or 1. The value of the logistic regression is between 0 and 1. Therefore it forms a curve like "S" which is called the sigmoid or the logistic function.

SVM: the aim of this algorithm is to create the best line or the decision boundary that can segregate the n-dimensional space into classes so that we can easily put the new data point in the correct category in the future this best decision boundary is called hyperplane. extreme points or vectors are chosen by the

SVM to create the hyperplane. The data points or the vectors that are close to the hyperplanes and which affects the position of the hyperplane are called support vectors.

C. Dataset

- From the UCI repository the dataset is taken. Allhypo.data and Allhyper.data has been combined to make the final dataset. The dataset has 7544 rows and 30 columns.
- The attributes like age, sex, TSH, T3, TT4, T4U, FTI, TBG are used.
- TSH is a thyroid stimulating hormone, T3 is a Triiodothyronine hormone that affects almost psychological process in the body, TT4 is a thyroxine hormone that evaluates the thyroid function and diagnose the thyroid disease, FTI is a thyroxine index that remains constant in healthy individuals. Hyperthyroid causes increased FTI.
- All the thyroid hormones are numerical values in the dataset.
- At the end we have the dependent variable i.e., the predict class that has labels like hyperthyroid or negative etc.
- Along with it the dataset has also some categorical values like sick, pregnant, I131 treatment (Iodine 131 treatment), thyroid surgery (to know if someone has undertaken thyroid surgery) etc.

Models	Accuracy
Logistic Regression	94.13
Support Vector Machine	94.13
Random Forest	93.85

From the above table we can see that the logistic regression and support vector machine gives the same accuracy i.e., 94.13% and we can consider them. As well as SVM, & Logistic Regression gives the highest accuracy as compared to the random forest.

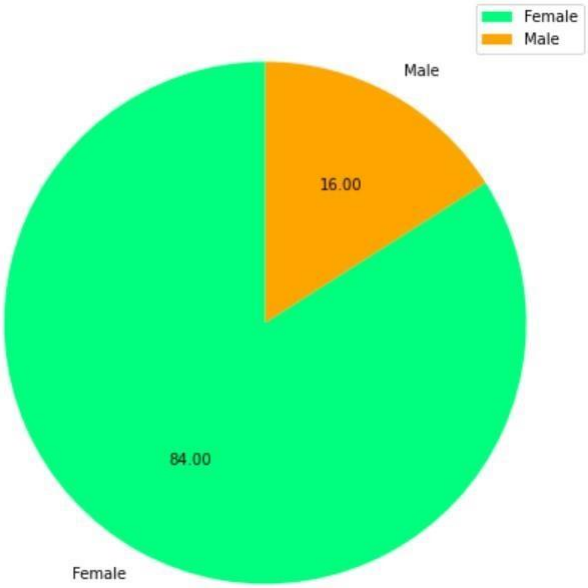
We also found that, if we compare the results of the logistic regression to the previous work our models give the highest accuracy, and we are also using different machine learning models.

V. CONCLUSION

IV. RESULT & ANALYSIS

After performing the data cleaning and EDA on the dataset we have come to know that 97.33% people are suffering from hyperthyroid and 2.67% people are not suffering from hyperthyroid.

Also, we have come to know that the number of female patients is more than the male patients. There are 84.00% of female patients and 16.00% male patients.



Below is the accuracy score of the models that has been done by other people.

Models	Accuracy
Logistic Regression	81.25
KNN	96.87
Decision Tree	87.5

The efficiency of the algorithm depends upon the dataset and the features that are selected for the prediction. Different models have different accuracy score during the evaluation. Below is the accuracy score of the models that we have achieved.

The intent of our work to be done further is to implement the ANN and deep learning algorithms on our dataset so that to achieve better prediction with the high accuracy score.

We have predicted the thyroid disease with better accuracy by using different machine learning models. Here the models are trained to detect whether the person is suffering from thyroid or not.

Different analysis has been done to predict the thyroid disease all with different accuracy score. A further better result can be achieved in the future if anyone can gather a better live dataset on thyroid and by applying deep learning concepts.

VI. REFERENCES

REFERENCES

[1] Thyroid Detection Using Machine learning" Published Online January 2021 in IJEAST.

[2] Interactive Thyroid Disease Prediction System Using Machine Learning Technique" by Ankita Tyagi, Ritika Mehra, Aditya Saxena published on 5th IEEE international conference on parallel, Distributed, and Grid Computing (PDGC-2018), 20-22 Dec.

[3] Thyroid Disease Prediction Using Machine Learning Approaches" published in the National Academy of Science, India 2020 done by Gyanendra Chaubey, Dhananjay Bisen, Siddharth Arjaria, Vibhash Yadav.

[4] https://www.javatpoint.com/machine-learning

[5] Dataset Reference: UCI Repository https://archive.ics.uci.edu/ml/datasets/thyroid+disease

[6] https://medlineplus.gov/labtests/triiodothyronine-t3-tests/ [7] A study of some data mining classification techniques in IRJET journal.

[8] Machine Learning Techniques for thyroid disease diagnosis – a review by Shaik Razia and M.R. Narasinga Rao published in Indian journal of science and technology.

[9] Disease prediction using machine learning.

[10] Diagnosis of various thyroid ailments using data mining classification techniques by Umar Sidiq, Dr.Syed Mutahar Aaqib, Dr.Rafi Ahmed khan. [11] Normal-thyroid-hormone-levels by Michael Yeh.