International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 08 | Aug 2022 www.irjet.net p-ISSN: 2395-0072

Comparative Analysis of Early Detection of Hypothyroidism using Machine Learning Techniques

Ranjitha B[1], K R Sumana[2]

[1]PG Student, The National Institute of Engineering, Mysuru, Karnataka, India
[2]Assistant Professor, Mysuru, Karnataka, India

Abstract - *The diagnosis of health conditions and proper treatment of disease at an early stage is one of the most challenging tasks in the healthcare field. Hypothyroidism is a type of thyroid disease. Thyroid glands are located in the middle of our necks. It has a butterfly shape and is small in size. People with hypothyroidism do not produce enough thyroid hormone to keep their bodies functioning normally. The thyroid gland may be involved in several conditions either directly or indirectly. Damage to the thyroid gland and inflammation are the causes of hypothyroidism. Low thyroid hormone levels cause the body's functions to slow down, leading to general symptoms like fatness, low pulse, increase in cold sensitiveness, neck swelling, dry skin, hands symptom, hair drawback, serious emission periods. The purpose of this project is to predict the Hypothyroidism disease at the early stage. Nowadays, machine learning has become an incredibly popular way to detect various diseases. Machine learning is used to detect disease at an early stage with greater accuracy. This Project uses KNN, Random Forest(RF) and XGB algorithms to predict the hypothyroidism disease at the early stage.*

Key Words: **Thyroid disease, Hypothyroidism, KNN, Random Forest ,XGB**

1.INTRODUCTION

Thyroid is one of our glands, which make hormones. Thyroid hormones control the rate of numerous conditioning in our body. It secretes a few chemicals that are blended in with blood and excursion across the body to control modling. There are two primary thyroid chemicals Triiodothyronine(T3) and Thyroxin(T4). These two chemicals are significantly answerable for keeping up with the energy in our bodies. The two main types of thyroid condition are Hypothyroidism and Hyperthyroidism. Hypothyroidism is caused when the gland releases low situations of thyroid hormone. Symptoms of an underactive thyroid(hypothyroidism) can include Feeling tired, Gaining weight, passing obliviousness, Having frequent and heavy menstrual ages, Having dry and coarse hair, Having a coarse voice. Foods that affect thyroid are tofu, tempeh, edamame sap, soy milk, etc. The potables coffee, green tea, and alcohol — these potables may irritate our thyroid gland. People who worked 53– 83 hours per week were shown to have a higher rate of hypothyroidism than those who worked 36–42 hours per week. The night shift work might be associated with the threat of subclinical hypothyroidism, and that this threat increased with longer employment as a night shift worker. While sleeping further than eight hours per day may increase the threat of both hyperactive and underactive thyroid function. Thyroid disease affect an estimated 200 million people worldwide. In India there are 42 Million people have thyroid diseases and Hypothyroidism is utmost of the common thyroid complaint in India.

2. RELATED WORK

The authors [1] in this article applied the classification (KNN) and prediction model (decision tree) to the thyroid dataset to accurately predict new patient entry. The KNN algorithm is used to classify thyroid disorders with related prioritized symptoms. Artificial Neural Network, support vector machine, Naive Bayes and KNearest Neighbor are the important modes applied to the prediction of thyroid disease and the results show that the K-nearest neighbor accuracy is better than any other thyroid disease detection technique. [2]utilized information mining calculations, for example, KNN, Naive Bayes, Support Vector Machine for the concentrate in this paper. The after effects of these arrangement techniques depend on the precision and execution of the model. For the given dataset, SVM accuracy is 0.82, Naive Bayes accuracy is 0.83 and KNN accuracy is 0.85. [3] Utilizes calculations like KNN, Random Forest, Naive Bayes, and ANN. KNN with Random Forest exhibited improved results with a precision of 94.8 percent when contrasted with the complete outcomes with four classifiers on the equivalent dataset. Utilizing decision tree algorithm, random forest algorithm, support vector machine algorithm, logistic regression and multilayer feedforward algorithm[4]. After doing a comparative analysis to identify the prediction algorithm that produces the most precise and accurate results, it can be said that the decision tree algorithm does so with a 99.46 percent accuracy rate and precision 0.99. The informational collections for the thyroid sicknesses have been had from the UCI website. The Machine Learning Algorithms like Artificial Neural Network, Support Vector Machine, Decision Tree, K-Nearest Neighbor are utilized to arrange and anticipate the exactness. Thyroid infection prescient models which require least number of boundaries of an individual to analyze thyroid illness and sets aside both cash and season of the patient. [5]This paper studies on thyroid disease and apply some algorithms to test performance study on mentioned algorithms. ANN-97.50,

KNN-98.62, SVM-99.63 and DT -75.76. Without a doubt, many thyroid diseases have been successfully diagnosed by experts worldwide. However, it is recommended that patients employ fewer diagnostic criteria when seeking a thyroid condition diagnosis. With more characteristics, a patient must undertake testing, which is both time- and money-consuming. In order to save patients' time and money, it is vital to develop algorithms and predictive models for thyroid disease that only require a few factors provided by the patient to detect the issue.

3. PROPOSED SYSTEM

Currently, machine learning has come an immensely popular medium for detecting various diseases. It's veritably accessible and effective to presume conditions using machine learning ways. We've used Machine learning algorithms such as KNN, Random Forest(RF) and XGB. Eventually derived that these algorithms helps us to attain better delicacy.

4. METHODOLOGY

Machine learning, a subfield of man-made brainpower, empowers PCs to "learn" for themselves from preparing information and work on over the long run without being expressly customized. Algorithms can make their own forecasts by recognizing designs in the information and gaining from them. Machine Learning Algorithms like Random Forest, K-Nearest Neighbor and XGBoost are utilized to anticipate the hypothyroidism in beginning phases.

4.1 ALGORITHMS

[1] Random Forest

A supervised learning algorithm is random forest. However, classification issues still account for a large portion of its use. As we all know, trees make up a forest, and more trees equal a more healthy forest. Additionally, the random forest algorithm builds decision trees from data samples, extracts forecasts from each one, and then uses voting to select the fashionable outcome. An ensemble system reduces over-fitting by combining the results, making it superior to a single decision tree.

[2] KNN

The KNN represents "K-Nearest Neighbor". The algorithm can be utilized to break both section and regression issue proclamations. The picture "K" stands for the number of nearest neighbours to another ambiguous variable that must be anticipated or organised. KNN operates by taking a chance on the distances between a question and each embodiment in the data, selecting the predetermined number representations (K) nearest to the inquiry, likewise votes in favor of the most successive data.

[3] XGB

Extreme Gradient Boosting, or XGBoost, is an idea put out by University of Washington specialists. Loads are critical in XGBoost. Every free factor is given a load prior to being taken care of into the choice tree that gauges results. Factors that the tree inaccurately anticipated are given more weight prior to being set into the subsequent choice tree. These unmistakable classifiers/indicators are then joined to deliver a hearty and precise model.

4. DATASET

The dataset are collected from the kaggle website. In this project we used some attributes to get patient details like Age of the patient, Sex-Patient Male/Female and some of the clinical details like Thyroxin, Antithyroid_medication,Goiter,Psych,T3, TT4, T4U, FTI.

TABLE -1: Dataset List

Attribute name	Description
Age	Age of the patient
Sex	Male/Female
Thyroxin	Clinical Test True/False
Antithyroid_medication	Clinical Test True/False
Goiter	Clinical Test True/False
Hypopituitary	Clinical Test True/False
psych	Clinical Test True/False
T3	Clinical Test Value
TT4	Clinical Test Value
T4U	Clinical Test Value
FTI	Clinical Test Value
Class	negative, compansated_hypothyroid, primary_hypothyroid, secondary_hypothyroid

5. SYSTEM ARCHITECTURE

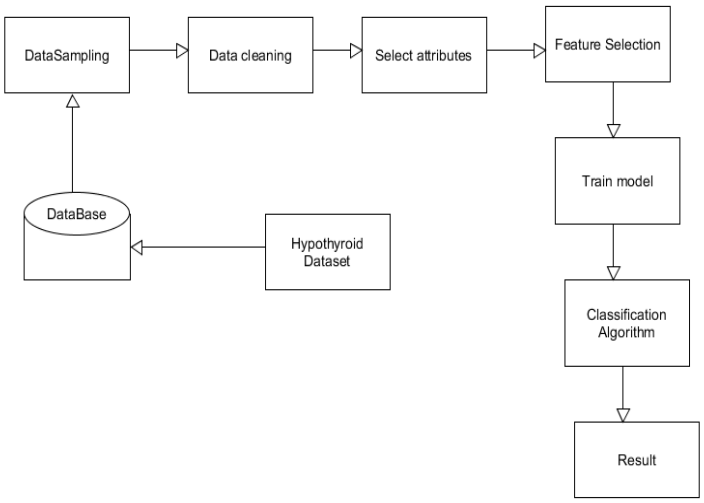

FIG -1: System architecture

Fig. No. 1 shows the system architecture, First collect the dataset from the website and dataset is stored in the database. Taken by the website are larger dataset that must be sample to balance. Then Clean the missing values and select the important attributes. Then apply feature selection and train the model. Then divide into testing and training model, apply algorithms i.e., knn, rf, xgb for getting better accuracy and finally we get result as early prediction of hypothyroidism.

6. EVALUATION AND TESTING

6.1 DATA PREPROCESSING

The data collected from the website are not formatted clearly. So, we should check for missing values in the list. Drop the unnecessary attributes and keep necessary attributes. Then taking care about catagorical data i.e., in our dataset-class attribute had negative, compensated_hypothyroid, primary_hypothyroid secondary_hypothyroid to find hypothythyroidism in early stages these classes are used. Then change object type to numerical type data. To apply machine learning algorithms, data must be split into training and testing.

6.2 TRAINING SET

To check training and testing, we should separate independent and dependent variables as X and Y and check for imbalanced data. By using Kmenas clustering will clusters the data. The class attributes are negative, compensated_hypothyroid primary_hypothyroid secondary_hypothyroid and the values as 0, 1, 2, 3 respectively. Then split into traing and testing data. Then create model as x_tarin,y_train, x_test and y_test.

Then, apply machine learning algorithms-A model called random forest is built using several decision trees. When constructing trees and breaking nodes, the model randomly selects samples from the training data points. Apply x tarin, y train, x test, and y test for the Random Forest Classifier to obtain accuracy for this algorithm.

Then, XGB works as load prior to being taken care of into the choice tree that gauges results. Factors that the tree inaccurately anticipated are given more weight prior to being set into the subsequent choice tree. These unmistakable classifiers/indicators are then joined to deliver a hearty and precise model and apply xgb model to x_tarin,y_train, x_test and y_test to get accuracy for this algorithm.

7. RESULT

After the implementation using KNN, Random forest and XGBoost Machine learning algorithms, and all of the classifier results are compared. Then evaluated the results based on the early stages like negative, compensated_hypothyroid, primary_hypothyroid, secondary_hypothyroid. Fig.No. 2 Shows the result of the hypothyroidism dataset. Classes are negative, compensated_hypothyroid, primary_hypothyroid, secondary_hypothyroid and the values are 0, 1, 2, 3 respectively. Here x-axis defines the class, in which class the patients are in highest count according to given dataset (negative, compensated_hypothyroid, primary_hypothyroid, secondary_hypothyroid) and y-axis defines the count of the patients(0, 500, ….3500) data. The dataset values are almost compensated_hypothyroid patients only indicates that early stage of hypothyroidism patients are found as shown in the chart-1 below and the accuracy for these algorithms are shown in table below.

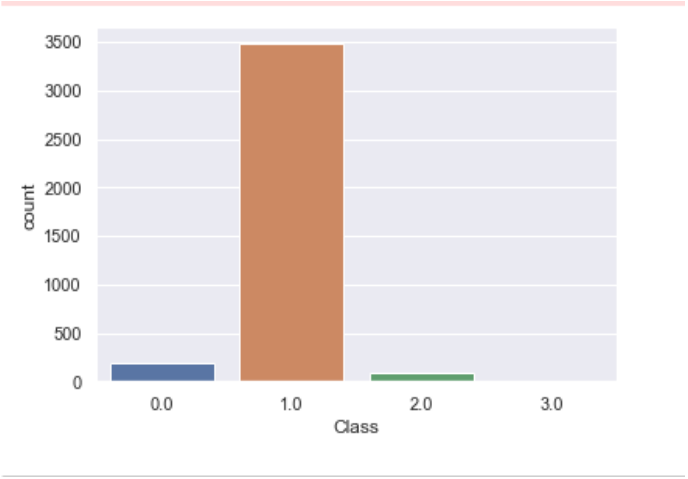

CHART -1: Number of Patient's having disease at early stage

TABLE -2 Comparison of Algorithms

Algorithms	Accuracy	Score	Time Efficiency
Random-Forest Classifier	88.5%	0.897	0.3ms
XGBoost Classifier	87.8%	0.902	1.5ms

8. CONCLUSIONS

In our discoveries, we have seen that KNN, Random Forest and XGBoost Algorithms are utilized to assists us with foreseeing hypothyroidism in the beginning phase by utilizing a continuous dataset. In the proposed framework its seen that an impediment of information to work with. At order to find a better solution and be better prepared to predict illness in its crucial stage, we will need to work with a larger dataset in the future. We also hope that more people from our country will be interested in dealing with this illness. Trust that will enable citizens of our nation to maintain a healthy society.

REFERENCES

[1] Prediction Of Thyroid Disease Based On Classification Using Hierarchical Structure By Charan R, Akash Yadav M, Aprameya N Katti, Mohith P Global Academy Of Technology, Bengaluru Karnataka 560098, India

[2] Thyroid Disease Prediction Using Feature Selection And Machine Learning Classifiers By Dr. Dayanand Jamkhandikar , Neethi Priya

[3] Empirical Method For Thyroid Disease Classification Using A Machine Learning Approach By Tahir Alyas , Muhammad Hamid , 2 Khalid Alissa, Tauqeer Faiz, Nadia Tabassum , And Aqeel Ahmad

[4] Thyroid Prediction Using Machine Learning Techniques By Sagar Raisinghani, Rahul Shamdasani , Mahima Motwani, Amit Bahreja , And Priya Raghavan Nair Lalitha Department Of Computer Engineering, University Of Mumbai, Vivekanand Education Society's Institute Of Technology, Mumbai, India

[5] Interactive Thyroid Disease Prediction System Using Machine Learning Technique By Ankita Tyagi ,Ritika Mehra, Computer Applications, Aditya Saxena, Computer Science And Engineering Dit University Dehradun, India