IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

THYROID PROBLEM DETECTION USING NAIVE BAYESIAN CLASSIFICATION

[1]Praveen Hugar, [2] G.Pranay Goud, [3] K.Vishnu Vardhan, [4] Mohammed Faizuddin

[1]Assistant Professor, [2]Student, [3]Student, [4]Student
[1]Department of Information Technology,
[1]JBIET, Hyderabad, India.

Abstract: Thyroid disease is a commonly occurring disease. Thyroid disease can result from conditions that cause over-or under-function of the thyroid gland. The overactive thyroid is known as Hyperthyroidism and the underactive thyroid is known as Hypothyroidism. A TSH test is done to find out if your thyroid gland is working the way it should. TSH stands for “Thyroid Stimulating Hormone” and the test measures how much of this hormone is in your blood. However TSH test is not enough to detect all the problems of Thyroid. Additional tests like T3 test and T4 test need to be carried out. A T3 test measures the blood level of the hormone T3 (triiodothyronine), and T4 test measures thyroxin. A Naive Bayes classifier is a probabilistic machine learning model that’s used for classification task by using Bayes theorem of probability. It is a supervised learning algorithm which makes predictions for new data using Bayes theorem. Naive Bayesian Classification is used to classify the given sample and detect the thyroid problem.

Index Terms - **Hyperthyroidism, Hypothyroidism, TSH, Supervised Learning, Naive Bayesian Classification.**

INTRODUCTION

One of the fourth leading diseases in India is thyroid disease which acts as a serious threat to the society as the change in family population, climate, urbanization, food increases the occurrence of change in thyroid hormone simulation which leads to thyroid diseases. The techniques available in the data analytics is a boon for the healthcare industry. The analysis helps in the accurate prediction of diseases, by creating the knowledge prediction model for the patients by analysing the patient’s history. This help in accurate decision making for the clinicians to diagnose the disease.

The untreated hypothyroidism can be a hazardous disease which even leads to death. Surprisingly there are certain factors like change in food style, environmental changes and a balanced iodized intake keeps the thyroid hormone secretion in control. According to the survey in India, 1 out of 10 adults are suffering from any one of the thyroid diseases. Moreover the philistinism, inadequacy of treatment facility are the main reasons for this wide spread. The early detection and prediction of thyroid deficiency will reduce the risks of survival. Many precision and diagnosis models have been designed for the prediction of thyroid diseases. Data analytics is yet another scientific access that enhances numerous algorithms, formulation and scientific approach to interpret the knowledge from the bigger data sources. Classification techniques process the data from the larger data sets and group the data instance for better results. Henceforth the classification algorithms and techniques are universally acceptable for the healthcare industry for prediction of diseases.

The thyroid gland is the biggest gland (butterfly shape) in the neck, whose function is to stimulate thyroid hormone, which results in effect on nearly all tissues of the body. The main function of the thyroid is to maintain the regulation of the body’s metabolism. In general, disorders of the thyroid gland can be classified into two classes including hyperthyroidism and hypothyroidism. Hyperthyroidism occurs when the thyroid gland produces too much hormone, the body uses energy faster than it should. While the hypothyroidism happens under the condition when the thyroid simulating hormones doesn’t produce enough hormones, which means the body uses energy slower than it should be. There are many different reasons why either of these conditions might develop. Now, it is said that about 30 million Indians have at least any one form of thyroid disease, and people of all ages and races can have the possibility to get thyroid disease. It is surprised to find the women are five to eight times more likely than men to get thyroid disease especially in highly altitude areas. The symptoms are complex but easily confused with other conditions as well, which make tougher diagnosis of thyroid disease to be difficult.

Naive Bayes is a statistical classification technique based on Bayes Theorem. The statement of Bayes Theorem is $P(A|B) = P(B|A)\ P(A)/P(B)$. Naive Bayes is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets. Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. Bayes classifier calculate the posterior probability for every class and for each observation. Then it classifies the new observation based on the class with the largest posterior probability. Naive Bayes classifier performs well even when the dataset is small. So, the Naive Bayes classifier predicts the thyroid disease with more accuracy.

LITERATURE SURVEY

Earlier, the use of computer was to build knowledge based clinical decision support system which uses knowledge from medical experts and transfers this knowledge into computer algorithms manually. This process is time consuming and really depends on medical experts opinions which may be subjective. To handle this problem, machine learning techniques have been developed to gain knowledge automatically from examples or raw data. Here, a weighted fuzzy rule-based clinical decision support system (CDSS) is presented for the diagnosis of disease, automatically obtaining knowledge from the patient's clinical data.

The clinical decision support system for the risk prediction of patient's disease consists of two phases: automated approach for the generation of weighted fuzzy rules and developing a fuzzy rule-based decision support system. In the first phase, it can be used the mining technique, attribute selection and attribute weight, age method to obtain the weighted fuzzy rules. Then, the fuzzy system is constructed in accordance with the weighted fuzzy rules and chosen attributes.

To detect thyroid disorders various image processing techniques have been used. Most image processing algorithms consists of pre-processing, segmentation, feature extraction, feature selection and classification. The image processing algorithm takes the MRI image of the thyroid as input. The pre-processing step is done to reduce the noise and improve the quality of the image. In segmentation step, the region containing abnormalities are detected and then the features are extracted in feature extraction and the best features are selected. The classification is done on the basis of selected features from the image.

Finally, the experimentation is carried out on the proposed system using the dataset created by collecting test levels details of patients and the performance of the system is compared with the neural network-based system utilizing accuracy, sensitivity and specificity. In this project, an attempt is made to utilize Naive Bayesian classification to classify the Thyroid problem.

PROPOSED SYSTEM

We propose to implement Thyroid disease detection using Naive Bayesian Classification. The implementation comprises of four modules. The first module deals with adding the patient's details and Test levels details to the DB. The second module deals with creating the comma separated value file from the DB. The third module deals with calculating the needed means and variances for the NB Classification. The final module deals with classifying the given sample, using Naive Bayesian classification.

The dataset is created by collecting the patient details and the test levels details and adding to the database. They are stored in two different tables in the database. Here we are using SQLite database as it provides better performance and portability. SQLite is file-based, the database consists of a single file on the disk, which makes it extremely portable and reliable.

All the records of test levels table from the database are stored in a Comma Separated Value (CSV) file and this file acts as a dataset. The attributes of the dataset are TSH, T3, T4 and the class label Thyroid. The attributes TSH, T3 and T4 contains the test values of the patients and the class label Thyroid shows whether thyroid is detected or not detected for each and every patient record containing TSH, T3 and T4 test values.

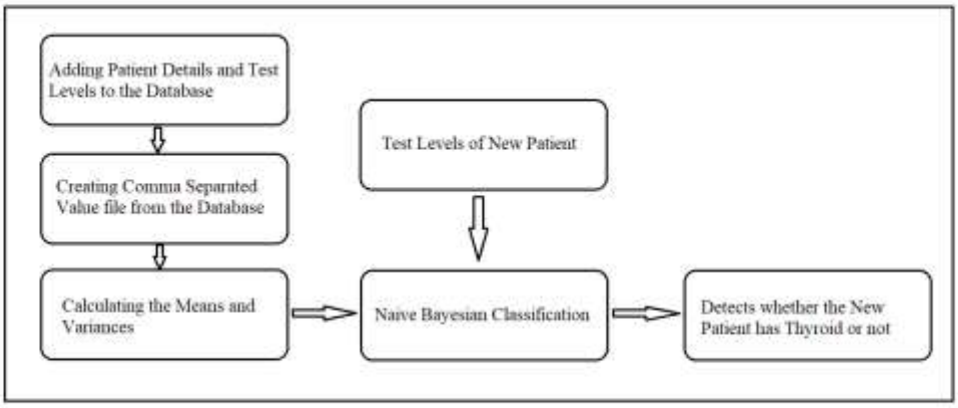
The means and variances of the test levels are calculated for detected and not detected thyroid cases separately. For classification purpose, we are using Naive Bayesian classification algorithm. This algorithm takes the new patient test levels details and the means and variances of test levels of the dataset as input. After taking the input, the algorithm classifies the new patient test levels and detects whether the patient has thyroid or not.

A user-friendly interface is designed for this project by using PyQt5 which is cross-platform python GUI toolkit. It provides Push buttons, Radio buttons, Check boxes and many more widgets. The user can enter patient's test levels in the text boxes provided in forms designed in PyQt5.

The designed form is saved as UI file. This UI file contains XML representation of widgets and their properties in the design. This design is translated into Python equivalent by using pyuic5 command line utility. The complete code for the project is written in Python as it contains numerous modules for dealing with machine learning algorithms and it provides more readability.

SYSTEM ARCHITECTURE

The first step in detection of thyroid disease is collection of patient's details and test levels details from various hospitals and adding them to the database. The dataset can be created by fetching the records from the database and storing them in a Comma Separated Value (CSV) file. From the dataset, we can calculate the means and variances of the test levels for both detected and not detected thyroid cases. These means and variances of the test levels are used for the prediction. A new sample of Patient test levels and the means and variances obtained from the database are given as an input to the Naive Bayesian classification algorithm. Naive Bayes algorithm applies Bayes theorem on the input and gives the predicted thyroid condition of the patient as output.


Adding Patient Details and Test Levels to the Database
Creating Comma Separated Value file from the Database
Calculating the Means and Variances
Test Levels of New Patient
Naive Bayesian Classification
Detects whether the New Patient has Thyroid or not

ALGORITHM:

NAIVE BAYESIAN CLASSIFICATION

Naive Bayesian classification algorithm is a probabilistic machine learning algorithm that classifies the input based on Bayes theorem of probability. Naive Bayesian classification is used for classifying many health care problems. In this project Naive Bayesian classification algorithm is used for detecting thyroid problem. Here, the algorithm takes the test levels of new patient as input and detects whether the patient has thyroid or not.

1. Importing library files. //Pandas and Numpy
2. Collecting Patient's details and Test levels.
3. Storing them in the database.
4. Creating the dataset by fetching values from the database.
5. Loading the dataset with attributes ['TSH','T3','T4'].
6. Calculating the Means and Variances of test values in the dataset.
7. Read test levels of the new patient.
8. Giving test levels of new patient to the Naive Bayesian classification.
9. Calculate the probability of thyroid using probability density function.
   P = 1/(sqrt(2*pi*variance_y)) * exp((-(x-mean_y)**2)/(2*variance_y))
10. Detected output showing whether the patient has thyroid or not based on probability.

MODULES

I. *Data Collection*

Firstly, Patient details and test levels details can be collected from various sources like health care organizations and hospitals. When the data collected is more, accuracy of the Naive Bayes classifier is also more. The data can be manipulated as per our requirement. Our data mainly consists of TSH, T3 and T4 test values. The collected data is added to the database. Here we are using SQLite database for storing patient details and test levels details.

II. *Dataset Creation*

After the collection of data, the test levels details of the patients are fetched from the database and stored in Comma Separated Value (CSV) file. We are using sqlite3 module to fetch data from the database and csv module for creating and writing test level records into the csv file. This csv file acts as the dataset. The attributes of the dataset are patient id, TSH, T3, T4 along with the class label Thyroid. All the attributes except the class label store continuous numeric values. The attribute Thyroid store categorical values which indicates whether the thyroid problem is detected or not.

III. *Calculation of Means and Variances*

The means and variances of TSH, T3 and T4 test values from the dataset are calculated for the detected and not detected thyroid cases. Here, we are using pandas module to read the dataset and to perform operations on the dataset. The mean method and var method of pandas module are used to get the mean and variance of values of a particular attribute or a record. Here, it generates the means and variances of the values of TSH, T3 and T4 attributes.

IV. *Classification*

Finally, the Naive Bayes classification algorithm uses probability density function to calculate the probability of thyroid. This function takes the new sample of test levels and the means and variances of test levels in the dataset as input and gives the prediction of thyroid disease by classifying the new sample.

CONCLUSION

This project is useful to analyse the Thyroid Hormone levels, using Naive Bayesian Classification. The project is useful to doctors in accurately classifying the Thyroid problems. It is also useful to the Thyroid patients, as in time measures can be taken, once any Thyroid problem is identified. This project finally leads to the improvement of quality of the patient's life. The dataset used in this project can be extended by adding more number of patient details and test levels details to the database. This increases the accuracy of the Naive Bayesian classification algorithm in detecting Thyroid diseases.

REFERENCES

[1] Bibi Amina Begum, Dr.Parkavi A- "Prediction of thyroid Disease Using Data Mining Techniques" in International Conference on Advanced Computing & Communication Systems (ICACCS), IEEE 2019.

[2] Roshan Banu D, and K.C.Sharmili "A Study of Data Mining Techniques to Detect Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 11), September 2017.

[3] Umadevi S, Dr.JeenMarseline K.S, "Applying Classification Algorithms to Predict Thyroid Disease" International Journal of Innovative Research in Science, Engineering and Technology (Vol. 6, Special Issue 10), September 2017.

[4] Polepogu Rajesh, Kunduru Umamaheswari, "Thyroid Disorder Detection Using Image Segmentation in Medical Images" International Journal of Scientific Development and Research (Vol. 1, Issue 6), June 2016.

[5] Ms.Wrushali Mendre, Dr.R.D.Raut, "Thyroid Disease Diagnosis using Image Processing: A Survey" International Journal of Scientific and Research Publications (Vol. 2, Issue 12), December 2012.

[6] PyQt5 [Online]. Available: https://riverbankcomputing.com/software/pyqt/intro [Accessed: 22-Feb-2020].

[7] SQLite [Online]. Available: https://www.tutorialspoint.com/sqlite/index.htm [Accessed: 22-Feb-2020].