THYROID DISEASE PREDICTION USING FEATURE SELECTION AND MACHINE LEARNING CLASSIFIERS

Dr. Dayanand Jamkhandikar 1, Neethi Priya2

1 Professor, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India.
2 Student, Computer Science and Engineering Department, GNDEC Bidar, Karnataka, India

ABSTRACT: Hypothyroidism or hyperthyroidism is a major disease in India which arises due to malfunctioning of thyroid hormones. In the traditional way diagnosis includes clinical examination and the many blood tests. Diagnosis of Thyroid Disease is very tedious and difficult tasks at early stages with high accuracy. Medical industry has enormous quantity of data, but the bulk of this data is not processed. For proper diagnosis data must be processed accurately. For accurate processing intelligent Machine learning techniques can be used. Machine learning algorithms have been employed to model the prediction and diagnosis of thyroid patients. In this paper an attempt is made to analyze naïve bayes, k-nearest neighbour and Support Vector Machine (SVM) for multiclass classification of thyroid dataset. With comparative study, different ML techniques will able to achieve better accuracy in disease prediction.

Keywords: hyperthyroidism, hypothyroidism, machine learning, feature selection, SVM, KNN and naive bayes

1. INTRODUCTION

The thyroid is a little gland in the neck that produces thyroid hormones. It may produce too much or too small of these hormones. Hypothyroidism is a situation in which thyroid gland is not able to produce sufficient thyroid hormones. These hormones regulate metabolism of the body and further affects how the body uses energy. Lacking the accurate amount of thyroid hormones, body’s normal functions start to slow down and body faces changes each day. In human services then clinical science, the applications based concerning Data dig are extremely gainful and significant. Analysis on Thyroid Disease is fairly dark or difficult errands. In medical field, Data mining assumes a quintessential assignment for finding concerning ailment. Information Mining gives several arrangement strategies in conformity with the forecast of ailment precision. The inert perception records gathered from plenty medicinal services association is treasured for the venture factors trial because some infections.

For diagnosis entire medical history and physical tests are used. As these tests produces large amount of data and ML can be used for finding important features from large amount of data. Due to this specialty of ML can be used in combination with medical science for the accurate diagnosis of hypothyroidism disease. A number of ML techniques have been evolved and in order to achieve best accuracy of a model ensembles are widely used. In this paper we are utilizing 3 calculations according to anticipate thyroid illness at starting period by Utilizing a number highlights ultimately we are waiting for the exactness about the result or looking at it.

1.1 Overview concerning thyroid

The thyroid is an endocrine organ as shown in below figure 1. The capability concerning thyroid member is in accordance with relinquish of thyroid hormones. It compasses in imitation of each single vile part thru the circulatory provision then control digestion or improvement. The extensive elements concerning thyroid part incorporate breath, blood flow, gut developments, temperature control, muscle working, assimilation and working concerning cerebrum. Any colorings action in the thyroid part might also affect the ordinary physiological work ethnic body. The thyroid hormone influences the development and improvement relying on the excuse on immunity. At the point now the creation over thyroid hormone is much less or is recognized so hypo-thyroidism. At the point so the introduction concerning thyroid hormone is high below it type on thyroid contamination is known as much hyper-thyroidism.

Fig. 1 thyroid organ
1.2 Thyroid and its fitness impacts
Thyroid difficulty are near basic endocrine infection, over the round the world. In an Indian study 42 million folks are experiencing these maladies. The Thyroid ailments are unique in relation in conformity with others as a long way as much their relative deceivability, utterance ease, scientific remedy mode availability .The incorrect advent over thyroid hormone influences wellbeing conditions.

1.2.1 Hyperthyroidism
Increment creation in the thyroid hormones causes hyperthyroidism Graves.' disorder is some of the immune system difficulty up to expectation causes hyperthyroidism .The warning signs are glacial pores and skin , increment affectability to heat ,diminishing regarding cloud ,weight reduction, raise pulse, hypertension, overabundance perspiring, neck extension, anxiety, menstrual intervals abbreviate, visit gut traits then fingers trembling .

1.2.2 Hypothyroidism
Reduction introduction among the thyroid hormones motives Hypothyroidism. The medical term hypo implies inadequate or less. The Symptoms incorporates corpulence, low pulse, and increase between tranquil affectability, neck expanding, dead skin, arms deadness, hair issue, violent menstrual intervals and stomach associated issues. What's more, it Symptoms may decompile upstairs period postulate not rewarded.

Thyroid hormones: The thyroid body produces are triiodothyronine (T3) or L-thyroxine (T4).The thyroid hormones manage distinct metabolic exercises, for example, age regarding warmth, the utilization over sugars, protein and fats. The pituitary organ controls advent about triiodothyronine and L-thyroxine hormones. The Thyrotropin-Stimulating Hormone from pituitary part is discharged now thyroid hormone is required then circles through the habit rule according to enter at thyroid organ. TSH at so much point animates the thyroid organs because the advent concerning T4 then T3 hormones. The advent concerning thyroid hormone is confined by way of the input arrangement on pituitary organ. The TSH creation is less then T3, T4 are greater of the dissemination and TSH advent is more when T3, T4 are less.

2. LITERATURE SURVEY
The author in [1] examined and then recommends the object over rarely any statistics dig tactics for method regarding thyroid sickness. Malady determination assumes a indispensable assignment and it is vital because of somebody top clinician. Thyroid sickness is one sizeable contamination and augur is the particularly troublesome assignment. Irina Ioniță yet LiviuIoniță” [2] and [3] suggested that the float discipline alludes after thyroid illness order into twins on the almost general thyroid dysfunctions (hyperthyroidism yet hypothyroidism) among the populace. The creators examined yet seemed at four characterization models [4] [5] Naive Bayes, Decision Tree, Multilayer Perception and Radial Basis Function Network. Ali keles et al. [6] proposed an expert system for predicting of thyroid that is known as Expert System for Thyroid Disease Diagnosis(ESTDD).This expert system diagnose thyroid diseases through neuron fuzzy rules with 95.33% of accuracy. S. B. Patel [7] worked to predict the diagnosis of heart disease patients using classification mining techniques. Three classification function techniques in data mining are compared for predicting heart disease with reduced number of attributes.

3. PROBLEM STATEMENT

3.1 Existing System:
Diagnosis of Thyroid Disease is very tedious and difficult tasks. The diagnosis thyroid disease in the traditional way includes clinical examination and the many blood tests. But then the main task is to diagnosis the disease at early stages with high accurate percentage. In medical field, Data mining plays a crucial role for diagnosis of disease. Data Mining provides many classification techniques for the prediction of disease accuracy. The need of patient data collected from much health care organization is useful for the risk factors analysis for many diseases. The clinical decisions are usually based on the doctor’s intuition. Therefore this may lead to disastrous consequences. Due to this there are many errors in the clinical decisions and it results in excessive medical costs.

3.2 Limitations of present fabric
- There is no action because health records yet ML techniques between existing explorations.
- No trial of the previous information.

3.3 Proposed System
In healthcare services data mining technique is mainly used for making decision, disease diagnosing and giving better treatment to the patients at comparatively low cost. Classification of thyroid disease plays an important task in the prediction of disease. Dimensionality reduction may be done as a future work so that number of blood test the thyroid will be reduced and also time required diagnosing disease. The thyroid Dataset is taken from UCI data repository site. The Database consists of thyroid patient records. The Patients record is having different attributes described in the data set.
description and different data mining techniques are applied to get the predication of thyroid disease. Data mining Algorithms such as KNN, Naïve bayes and Support vector machine are considered for the study.

3.4 Advantages of proposed system:
- We can predict the results using best classifier.
- Dynamic nature in prediction.
- We can predict on our own by collecting the readings from clinical test.

4. SYSTEM ARCHITECTURE

![System Architecture Diagram]

The above figure 2 shows the system architecture as the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. First of all we collect the different patient's data. Out of many attributes we only select 15 attributes through feature selection. After extracting the features we apply the classification algorithms SVM, KNN and naïve bayes. Admin will train the data and can predict the best classification algorithm based on the accuracy of the result.

5. IMPLEMENTATION

Machine learning employs three main different algorithms those are explained below:

5.1 Support Vector Machine (SVM)
One on the kind of lesson fabric tab is Support Vector Machine[8] ,which is utilized in accordance with function characterization among a excellent exactly and utilizations 2 class classifier, alluded as much constrained airplane as much "choice power then choice surface". The atypical plane isolates fine preparing take a look at along the bad making ready facts check within an arrangement. The points over pastime comprises a simple expand, utilized because of graph rearrangement, quadratic enchantment issue execute remain defined.

5.2 k-Nearest near (KNN)
KNN [9] is some about the best regular AI calculations utilized because order, It companies an information point structured of whether its neighbors are characterized. KNN stores every alone reachable litigation and organizations a density measures. The k-Nearest Neighbors tab (or KNN because short) is an easy tab in accordance with comprehend yet in accordance with execute. The usage choice keep explicit because of association troubles yet wish stand shown utilizing the Iris blossoms characterization issue. The model because of KNN is the whole preparing dataset. At the point then a port end is required because of a concealed facts example, the KNN calculation choice seem via the coaching dataset because of the k-most comparative cases. The augur faith over the almost comparable occasions is summed above then lower back so the hope because of the inconspicuous occurrence. The resemblance measure is reliant on the type about information. For authentic esteemed information, the Euclidean severance be able be utilized. Different kinds about information, because example, complete then doubled information, hamming split may stay utilized. On account over relapse issues, the regular of the predicted tension may stay returned.

5.3 Naive Bayes (NB)
In AI naïve bayes [10] classifiers are a crew regarding straightforward probabilistic classifiers dependent about applying bayes hypothesis with sure autonomy presumptions of the highlights. They are among the least complicated Bayesian provision models. it shares a usual government as the proximity about a precise thing in category lamely after the proximity of partial mean element.

5.4 The dataset description:
Dataset is committed from UCI AI storehouse [11]. Database contains concerning sufferers thyroid records. Every thyroid patient's document is comprises of 15 characteristics files beneath. Characteristic execute remain Boolean (genuine/bogus) then steady esteemed are addicted beneath. Below figure 3 shows the data set Hypothyroid.csv. Figure 4 shows the dialog flow and figure 5 shows the user registration.
5.5 MODULES

ADMIN: Admin is the one who decides which classification algorithm has to be given to the user. Admin after login, train the dataset with accordant iii calculations. Support Vector Machine (SVM), k-Nearest Neighbor (KNN) and naïve Bayes (NB). The classification can be done by the admin. The following functions are performed by the admin:

- After the training admin will test the accuracy by splitting 30 percentages of data from the training file.
- Then admin will find the best classification algorithm, called Naïve bayes.
- Admin also can see the graph of the accuracy of the three algorithms and feature selection score graph.

USER: User is an end consumer of the application; our application intention help in imitation of the consumer by means of hope thyroid contamination by instruct the past patient’s dataset together with classification algorithms. User may login with his medical details to known whether he is suffering from thyroid disease or not. The functions of the user are given below:

- User can register with own details and after login user upload single patient record in the csv file.
- User can see the result with through the forecast of calculation.

6. RESULT ANALYSIS

Below figure 7 shows the classification window and figure 8 shows the graph of feature selection. The accuracy prediction for various machine learning approaches is shown in figure 9.
The graph of accuracy analysis using SVM, Naïve Bayes algorithm and KNN is shown in figure 10. Figure 11 identifies whether a person is having thyroid or not by displaying positive if the person is suffering from thyroid or else negative.

CONCLUSION

In this paper we have proposed method to predict the thyroid disorder at earlier stage using data mining techniques. Data mining classification algorithms are used to diagnose the thyroid problems. Proposed technique helps to minimize the noisy data of a patient. Data mining Algorithms such as KNN, Naïve bayes, Support vector machine are considered for the study. The results of these classification methods are based on accuracy and performance of the model. The resulting classification of effective data helps to find the treatment to the thyroid patients with better cost and facilitate the management. For the given data set the accuracy using SVM is 0.82, Naïve Bayes is 0.83 and KNN is 0.85.
REFERENCES

[1] Roshan Banu D, then K.C.Sharmili “A Study of Data Mining Techniques after Detect Thyroid Disease” International Journal concerning Innovative Research into Science, Engineering and Technology (Vol. 6, Special Issue 11, September 2017
[2] IrinaloniNă or LiviuIoniNă” Prediction on Thyroid Disease Using Data Mining Techniques”The Classification Technique because of Talent Management the use of SVM, (ICCEET), 978-1-4673-0210-4/12, pp. 959-963, 2017
[6] Ali keles et al. “Expert System for Thyroid Disease Diagnosis(ESTDD)”.This expert system diagnose thyroid diseases through neuron fuzzy rules with 95.33% of accuracy.
[7] S. B. Patel “Review of machine learning techniques ”also worked to predict the diagnosis of heart disease patients using classification mining techniques. Three classification function techniques in data mining are compared for predicting heart disease with reduced number of attributes.
[8] Umadevi S , Dr.Jeen Marseline 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
[10] Lewis, D. “Naive Bayes at forty” the independence assumption in information retrieval. It describes about the working of naïve bayes algorithm.