

AN ANALYSIS OF PREDICTIVE MODELS FOR THYROID DISEASE USING MACHINE LEARNING TECHNIQUES

Suresh Kumar Kashyap¹, Dr. Neelam Sahu²

¹Research Scholar Ph. D(CS), Dept. of IT & CS, Dr. C.V. Raman University Kota, Bilaspur (CG) India

²Associate Professor Dept. of IT & CS, Dr. C.V. Raman University Kota, Bilaspur (CG) India.

Abstract: Thyroid disease is now very common worldwide. In India too, there is a very high burden of thyroid diseases. According to various studies on thyroid disease, it is estimated that about 42 million people in India suffer from thyroid diseases. The main objective of this paper is to build prediction modeling of the given medical data of patients with and without thyroid. Through this paper, we aim to create hybrid models can be easily used by doctors to treat patients with thyroid. Naïve Bayes and Random forest algorithms are used to predict whether a person having thyroid or not, by keeping his health conditions in mind. Thus this process enables doctors to easily group, classify and categorize the disease type accordingly treatment can be given to them. It has been concluded that random forest algorithms gives better accuracy than the other algorithms.

Keywords: Classification, Thyroid Disease, Naive Bayes, Hypothyroidism, Accuracy, Data sets.

1. INTRODUCTION

The thyroid gland leads to various infections such as allergies, vascular diseases, infertility, and causes tumors in some cases. Making a step towards formal discovery and basic treatment is fundamental. Female patients face a higher risk of thyroid disease than men. A pregnant and teenage girl is also suffering from thyroid disease. The endocrine structure and certain endocrine organs, which have the thyroid, are equally tolerant of further organ structures and important changes in use in maturation. The various acceptable and physiological changes of the thyroid during the growth process are noteworthy. The current study focuses on timely findings related to changes in fluid volume during the path to growth. The thyroid gland often creates symptoms in the neck area where the thyroid is located and will swallow when the patient is presented with thyroid disease. The most common thyroid sections are Goiter (an enlarged thyroid gland). The thyroid gland releases more hormones than the human body is forced to develop Hyperthyroidism, when the thyroid gland does not produce enough thyroid hormones to cause Hypothyroidism, Thyroid growth, Thyroid knobs are factors that affect the thyroid gland and Thyroiditis is an inflammation of the thyroid gland only organ. Common symptoms of hypertension include heart rate, abnormal blood pressure, abnormal body temperature, swelling of the hands and feet, hair loss, obesity or unexpected weight loss, depression, mood swings, metabolism, memory loss, dry skin, itchy skin, sensitivity to the eyes etc. In female patients suffering from thyroid disorders their menstrual cycle can be abnormal or difficult to menstruate. During pregnancy, female thyroid patients have failed to help with reproduction, pregnancy, depression and breastfeeding. Women between the ages of 40 and 50 who experience menopause, menopause begins or have severe menstrual symptoms.

2. LITERATURE REVIEW

Predicting using a traditional model for a serious illness in general including machine learning and supervised learning an algorithm that uses training data with labels of modeling training. High-risk and low-risk patient the classification is made into groups for group testing. But these types they are only relevant to medical conditions and are widespread he learned. A sustainable health monitoring system using smart clothes by Chen et.al. He read well systems are different and he was able to achieve the best the effects of cost reduction on the tree and in the simplest way cases of various systems.

Mohammed Abdul Khaleel, Sateesh Kumar Pradhan, G.N Dash (2013). "A survey of data mining techniques for finding locally frequent diseases" This paper focuses on mining the required medical data to find frequently occurring diseases, such as breast cancer, heart illness, lung cancer and so on. Data mining techniques like Apriori and FPGrowth, linear genetic programming, decision tree algorithms, unsupervised neural networks, outlier prediction techniques, classification algorithm, NaïveBayesian and so on have been applied.

K. Vembandasamy, R. Sasipriya, E. Deepa (2015). "Aims on analyzing heart diseases using a Naïve Bayesian algorithm". The algorithm used here is Naïve Bayes, which firmly assumes that the presence of any attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent. The tools used are WEKA and classification is done by splitting data into 70% of the percentage split. The naïve Bayes technique used was able to produce 86.41% of the input data correctly and 13.58% of inaccurate instances. He uses a dataset collected from a leading diabetic research institute in Chennai which has about 500 instances or patients.

TawfikSaeed Zeki et al. "An expert system for diagnosing diabetics". They proposed rule-based IF-THEN system. They have used three modules for 3 stages, they are Block Diagram, Mockler Charts, and Decision Tables. After considering many factors, this system provides a diagnosis of diabetics. It was developed inVP-Expert.

Vishali Bhandari and Rajeev Kumar, "Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis " compared different fuzzy expert systems by using multiple parameters for diagnosing the diabetics. MATLAB fuzzy logic toolbox was used for the comparative study of these expert systems. Five parameters were used for comparison and results were generated.

3. METHODOLOGY USED

This section describes the algorithm, language and software used for this function. A set of data used for experimental purposes was downloaded from the site of the University of California at Irvin (UCI) (web source <http://www.archive.ics.uci.edu/ml/datasets.html>). The over all procedure for the thyroid gland diagnostic procedure is shown in Figure 1

DATA SET DESCRIPTION

Data Set Description:-I am use data set from UCI machine learning Repository that was used for implementation with 3772 instances of 23 independent attribute and 1 dependent attribute. The details of data set is show in following table

Thyroid Disease data set Attribute description:- We can see all attributes with value type in this table

S.NO.	Attribute Name	Value Type	S.NO.	Attribute Name	Value Type
1	Age	continuous,?.	13	goitre	f,t.
2	Sex	M,F,?.	14	TSH_measured	f,t.
3	on_thyroxine	f,t.	15	TSH- Thyroid Stimulating Hormone	continuous,?.
4	query_on_thyroxine	f,t.	16	T3_measured	f,t.
5	on_antithyroid_medication	f,t.	17	T3 - Total Triiodothyroxine	continuous,?.
6	thyroid_surgery	f,t.	18	TT4_measured	f,t.
7	query_hypothyroid	f,t.	19	TT4- Total Thyroxine	continuous,?.
8	query_hyperthyroid	f,t.	20	T4U_measured	f,t.
9	pregnant	f,t.	21	T4U	continuous,?.

10	sick	f,t.	22	FTI_measured	f,t.
11	tumor	f,t.	23	FTI - Free Thyroxine Index	continuous,?.
12	lithium	f,t.			

Table 1: Thyroid Disease data set Attribute description

My research work is divided in many stage that are following;-

1. Data preprocess
2. Features Selection
3. Splitting of Data set
4. Performance Evaluation
5. Result comparison

This diagram show the all stages of my research work.

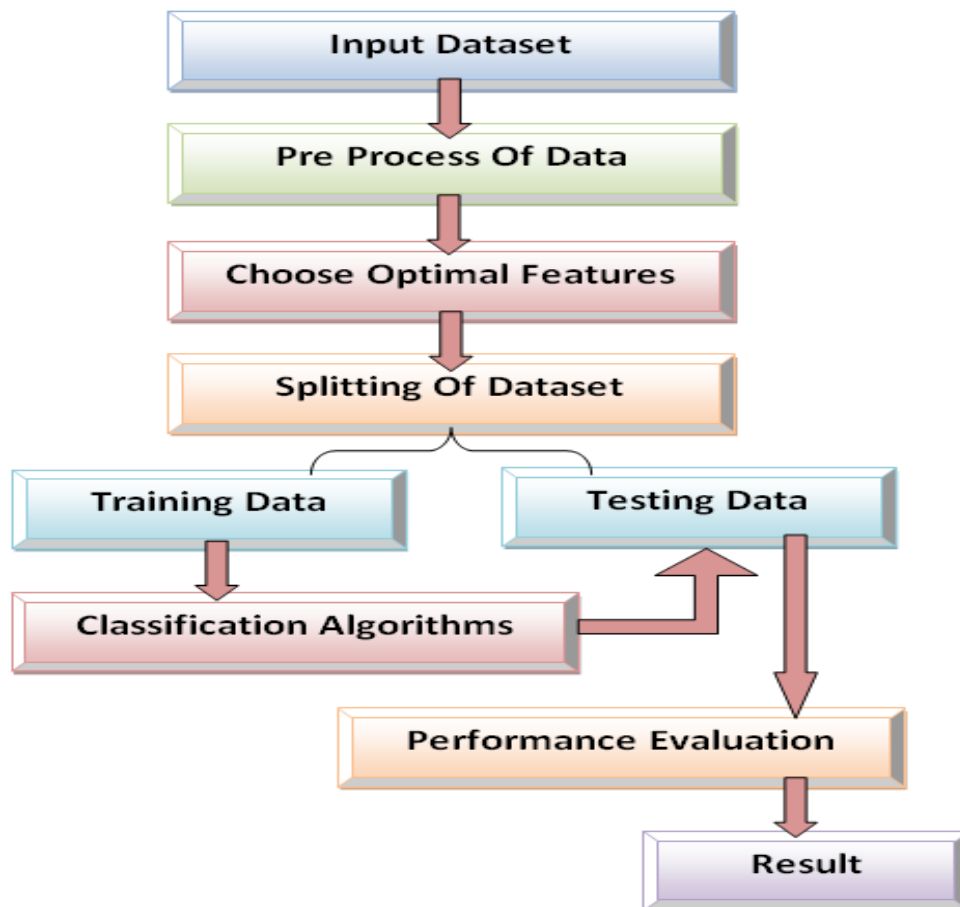


Figure 1 describes the flow diagram of the Prediction Process.

Naive Bayes

The Naive Bayes classifiers are a group of sophisticated algorithms supported by the Bayes' Theorem. It's not one algorithm but a family of algorithms within which all of them share the identical goal, which implies that each one of the components are independent of every other. The Naive Bayes separator is an example of the equipment accustomed to the separating function. The crux of separation relies on the concept of the Bayes theorem.

Bayes Theorem:

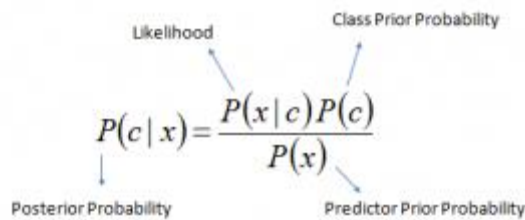
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

By using the Bayes theorem, we will detect the chance that A may have occurred, because B has occurred. Here, B is evidence and A is hypothesis. The belief made here is that the predictions / features are independent. That the presence of 1 element doesn't affect the opposite. Therefore it's called naive.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability $P(c|x)$ from $P(c)$, $P(x)$ and $P(x|c)$. Look at the equation below:



$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Above,

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of class.
- $P(x/c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of predictor.

Random Forests algorithm

Random Forests algorithm is a very powerful separation algorithm that can distinguish large amounts of data with high accuracy. The Random Forest is a group learning method (it is a method of close proximity neighbor) of divisions and regimes that forms the number of decision trees during training and ultimately the category with the largest number of votes will be considered as a predictor release. The variability of the tree forecast in each tree depends on the random vector values independently and the uniform distribution of all the trees in the forest. Random Forests solve this problem of high diversity and high bias by finding the natural balance between these two extremes. They also have a way of estimating error rates

(Without bag error). Many machine learning models, such as direct and systematic retrieval, are easily affected by vendors in training data. Outliers are changes in system behavior and can be caused by human error, tool error There is a chance that the sample provided will be contaminated. These exports or excess prices do not affect model performance / accuracy. RF Algorithm overcomes and solves this problem. In our case, we've split the labels into two variables, and this is for the input separator. One of the great strengths of the Random Forest categories is the ability to use any type of data, especially with feature selection. In our case, we use the RandomForestClassifier () function in the sklearn library to make data predictions. Input education and test data are modeled using fit (). Training details are then divided by the same members at the time of prediction. Model accuracy is obtained by comparing predicted values against the first set of values

4. RESULT:-

First, the accuracy of the training data was assessed by feed disputes over the separation of training data. After that, The accuracy of the test data is done in the same way as test data as parameters. By comparing these two, we can create a confusion matrix. The main purpose of confusion matrix to check the accuracy of separation. By definition a confusion matrix C is that, here $C_{i,j}$ indicate the number of observations known to be in group i but predicted to be in group j, here the count of true negatives are $C_{0,0}$, false negatives are $C_{1,0}$, true positives are $C_{1,1}$ and false positives are $C_{0,1}$

After performing Random Forest with Naive Bayes algorithms, we create the following results for different variations of training and assessment data: In table 2, we see that in four different divisions, we find results close to 75% in the training set and 74-77% in test results in Naïve Bayes section. This indicates that the training set has been trained to 75% accuracy which means that trained data is available used to predict 75% test results moderate accuracy in database analysis. In Random Forest section we can see that for the four different splits, we get results that are close to 98% in the training set and 72-77% in the test results As we analyze these table, we can understand that The random forest algorithm has a better training set result it also provides better prediction and analysis. The data set is trained for high accuracy where everything variables are assumed to be inserted without inserting missing data as the Random Forest algorithm will make sure it does not data not in big databases.

Train	Test	Prediction using Naïve Bayes		Prediction using Random Forest		
		Train Result (%)	Test Result (%)	Train (%)	Result	Test Result (%)
60	40	76.27	77.85	97.72		76.89
70	30	75.11	75.12	98.80		74.27
75	25	75.77	74.65	98.73		73.16
80	20	75.81	77.83	98.43		75.37

Table 2: Comparison of Training results for various splits

5. CONCLUSION

This paper presents a comparative study on thyroid disease diagnosis by using random forest classifiers and, Naïve Bayes . The results were compared and it was seen that Random forest classifiers is better than naïve bayes . Random forest classifiers could be successfully used to help the diagnosis of thyroid disease. It is observed that the Naïve Bayes not provide more accuracy for training data and test data compare than random Forest classifiers.

REFERENCES

1. Bichler M, Kiss C (2004) A comparison of logistic regression, k -nearest neighbor, and decision tree induction for campaign management. In: Proceedings of the tenth Americas conference on information systems, New York.
2. Geetha G, K.Mohana Prasad(2020),Prediction of Diabetics using Machine Learning International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-5.
3. Gorade SM, Deo A, Purohit P (2017) A study of some datamining classificatio n technique. Int Res J Eng Technol4(4):3112-3115.
4. <http://archive.ics.uci.edu/ml/mac hine-learning-databases/thyroid-disease>.
5. Lonita I, Ionita L (2016) Prediction of thyroid disease using data mining techniques. Broad Res Artif Intell Neurosci 7(3):115-124.
6. Mirza Shuja, Mittal Sonu, Zaman, Majid.(2018), Design and Implementation of Predictive Model For Prognosis of Diabetes Using Data Mining Techniques. International Journal of Advanced Computer Research, Vol. 9, Issue 2. pp. 393-398.
7. Monaco Fabrizio (2003) Classification of thyroid diseases: sug-gestions for a revision. J Clin Endocrinol Metab 88:1428-1432.
8. Patel Hetal. (2019) An Experimental Study of Applying Machine Learning in Prediction of Thyroid Disease. International Journal of Computer Sciences and Engineering, Vol. 7, Issue 1, pp. 130-133.
9. Roshan Banu D, K.C, Sharmili.(2017) 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. pp. 549-553.
10. Shrivasa, A. K. Ambastha, Pallavi.(2017) An Ensemble Approach for Classification of Thyroid Disease with Feature Optimization. International Education and Research Journal, Vol. 3, Issue 5. pp. 112-113.
11. Tyagi Ankita, Mehra Ritika.(2018) Interactive Thyroid Disease Prediction System Using Machine Learning Technique. 5th IEEE International Conference on Parallel, Distributed and Grid Computing, pp. 689-693.
12. Ulutagay G (2012) Modeling of thyroid disease: a fuzzy inference system approach. Wulfenia J 19(1):346-357.