Thyroid Prediction Using Machine Learning Techniques

Sagar Raisinghani[1](✉), Rahul Shamdasani[1], Mahima Motwani[1], Amit Bahreja[1], and Priya Raghavan Nair Lalitha[2]

[1] Department of Computer Engineering, University of Mumbai,
Vivekanand Education Society's Institute of Technology, Mumbai, India
{2015sagar.raisinghani, 2015rahul.shamdasani, 2015mahima.motwani, 2015amit.bahreja}@ves.ac.in

[2] Computer Engineering Department,
Vivekanand Education Society's Institute of Technology, Mumbai, India
priya.rl@ves.ac.in

Abstract. Thyroid is a critical medical condition which can be caused either due to increased levels of TSH (Thyroid Stimulating Organ) or due to some infection in thyroid organs itself. The machine learning algorithms have been employed to model the prediction and diagnosis of thyroid patients. A variety of these algorithms including Decision trees, Random forest, Support vector machine, Artificial Neural Network and Logistic regression have been widely used in development of predictive models of thyroid disease. The paper presents a review of recent ML algorithms applied in the prediction and diagnosis of thyroid detection. The proposed system is used for thyroid disease prediction of patients, based on various symptoms and reports of thyroid. With comparative study, different ML techniques are used by the proposed system to achieve better accuracy in disease prediction. Among these, Decision tree algorithm is found to be better with the accuracy of 99.46%.

Keywords: Machine learning · Predictive models · Thyroid prediction · Thyroid diagnosis · Thyroid classification

1 Introduction

According to the survey [9], it states that on an average one out of 38000 people in the world are suffering from congenital hypothyroidism. In developing countries like India, there are almost 42 million people suffering from thyroid disease. It seems to be more common among Indians, especially this ratio in Mumbai is stated as one out of 2640. Now a days more than 25,000 hospitals across the globe collects data on patients in various formats. In the traditional method, the clinical and medical studies are carried out using classical analysis and statistical tests.

Thyroid disease is widely spread in today's world and it often causes severe damage to life and body. It affects functioning of thyroid gland which in turn results into excess secretion of thyroid hormones. Its symptoms include low energy, weight gain, fatigue, inability to tolerate cold, dry skin, slow heart rate, there may be a swelling

M. Singh et al. (Eds.): ICACDS 2019, CCIS 1045, pp. 140–150, 2019.
https://doi.org/10.1007/978-981-13-9939-8_13

in a part of neck. In this disease body goes into auto safe mode in which hormones are generated which pulverize the thyroid organs. It affects the body in an irreversible manner so it is very important to avoid this disease. The avoidance of this disease requires preliminary knowledge of the occurrence of this disease as it is very difficult to cure this disease once it reaches its final stages.

2 Literature Survey

The research paper by Rao and Razia [1] showcased almost all the ML techniques with their basic structure. Out of all the algorithms, ANN showed best result. The major limitation of this method was it didn't explored the Genetic Algorithm for better optimized result.

Prerana et al. [2] proposed the technique of data mining using neural networks that can be used for early prediction of thyroid. The network was trained using back propagation and gradient method working simultaneously. But the variation in layers of various network parameters were not considered during training.

Umadevi et al. [3] worked on the trial of 21 parameters to train the model using classification algorithm. The model was trained using KNN, ANN and fuzzy ANN algorithm and the accuracy was compared. It was evident that Fuzzy ANN performed better than other two classification algorithms. But it is observed that due to few samples for dis-functions of thyroid, the classification between over and under functioning thyroid was difficult.

Ammulu and Venugopal [4] have proposed data mining technique to predict hypothyroidism in the patient. In the research paper, data mining technique is applied on the hypothyroid dataset to determine the positive and the negative cases from the entire dataset. But, the algorithm works only for under functioning thyroid and thus nothing can be said about hyperthyroidism.

Ahmed et al. [5] provides Support vector machine (multi, binary) algorithm for thyroid prediction. The precision value along with confusion matrix was used for evaluation of results. Medical data cleaning was used for filling all the blank spaces. When thyroid disease goes through structural changes it becomes difficult to detect it based on variations of thyroid hormones.

Mahajan et al. [6] have provided a way to detect hypo and hyperthyroid from thermal images using Bayesian classifier. It provided 81.18% accuracy in classification. The major drawback of this algorithm was if the image were not cleared or not processed properly, the results shown differed with a high range than the actual output.

Saiti et al. [7] have proposed thyroid prediction using PNN and support vector machine. Feature selection was done using genetic algorithms. The fitness evaluation contained two terms: (1) accuracy and (2) the number of features selected. When the algorithms were tested without the GA, the accuracy of both the methods was around 85%, but with the help of the GA accuracy obtained was nearly 100%.

The proposed system will even help the new practitioners to improve their analysis skills and predict the disease even without prior knowledge about it. The primary task is to provide thyroid diagnosis at early stages and also attain higher accuracy.

3 Methodology

The patient data provided can be either in structured or unstructured format. The unstructured data is required to be transformed into structured data in order to analyze the data.

3.1 Structured Data

Different approaches are discussed as follows for the better use of structured patient data in health prediction system:

Machine Learning Algorithms. Artificial neural network being an application of machine learning is a very powerful algorithm. It can work on structured data and it is always able to derive pattern from a dataset of any size.

A large number of attributes are considered in the dataset and considering the attributes and the vividness of the data we have shortlisted five algorithms for implementation out of which the best will be selected.

Decision Tree. It is one of the most important classification and prediction method in supervised learning. A decision tree classifier has a tree type structure which provides stability and high accuracy. Decision tree is also referred to as CART (Classification And Regression Trees). Decision tree applies simple if else rules to construct the trees. Decision tree algorithm commonly uses gini index, information gain, chi-square, and reduction in variance to make a strategic split.

Multilayer Feed Forward Neural Network. In neural networks the nodes are initialized these nodes are connected with each other. The line connecting them have a predefined weight, for many years there was no optimized technique to decide these weights, but now these weights are considered to be affecting the prediction results so they need to be optimized for which an iterative method is used which optimizes the results.

Support Vector Machine (SVM). It is a very useful algorithm sometimes in case of large data it gives very good efficiency it is used for classification purpose, it classifies the data into various sets and trains the model using this set and then it successfully predicts the data using this trained model. First it plots the attributes in the graph and separates it broadly using a boundary and the nodes which lie on this boundary are named as the support vectors.

Logistic Regression. Logistic regression conducts regression analysis when the dependent variable has dichotomous results (binary). Logistic regression is a very strong technique for doing predictive analysis. It defines the data very efficiently and explains the relation between one dependent binary variable and one or more nominal independent variable. One of the important consideration done while selecting the logistic regression model is the model fit. Selection of variables is another important because more the number of scrap variable lesser will be efficiency of the algorithm.

Random Forest. Random Forest is a supervised learning algorithm. The general idea behind the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Random Forest has almost the same hyper parameters as a decision tree or a bagging classifier.

4 Proposed Model

While training the model, various machine learning algorithms are evaluated and the best is selected. This in turn will help in increasing the efficiency of the system. The attributes which are selected are taken from the health expertise.

4.1 System Architecture

The Fig. 1 below shows the conceptual design of the system that is being proposed. Components of the proposed system are described as following:

Data Pre-processing. The data obtained must be preprocessed into an understandable format. In order to preprocess the data, check out the missing values if any in the dataset. If there are some missing values then they must be replaced with mean, medium or mode of the feature. Then, the categorical data is required to be transformed into numerical data. To apply machine learning algorithm on the dataset, Dataset is split in training and testing set.

Training Set. The training data is then trained using a machine learning algorithm. In decision tree algorithm, all the attributes are tested for the split using cost function, gini index in our case. A root node is obtained after the first split. A higher gini index indicates greater inequality, and thus the split occurs at the attribute which has the least gini index value.

In Support Vector Machine algorithm, the numeric input variables in the thyroid dataset (the columns) form an n-dimensional space. The right hyper-plane is identified that divides the two classes, either class 0 or class 1.

In Random Forest algorithm, a decision tree is an intuitive model and the building block of this algorithm. But, the decision tree algorithm is prone to overfitting as the maximum depth is not limited. Random forest is a model constructed with many decision trees. The model randomly samples the training data points when building trees and considers random subsets of features when splitting nodes.

In logistic regression algorithm, the value of dependent variable is predicted using independent variables. It is assumed that there is a relation between dependent variable and predicted variable. To train the data, coefficients are found that best describes the predictor variables for the linear relation.

Data Pruning. In decision tree algorithm, pruning is done to improve the performance and stability of the tree. The complexity if tree is reduced by removing the less important branches of the tree. Pruning increases the accuracy of the algorithm and also reduce overfitting.

In Support Vector Machine algorithm, reduce the complexity of the function and increase the speed of SVM. Iterative process is used to prune SVM.

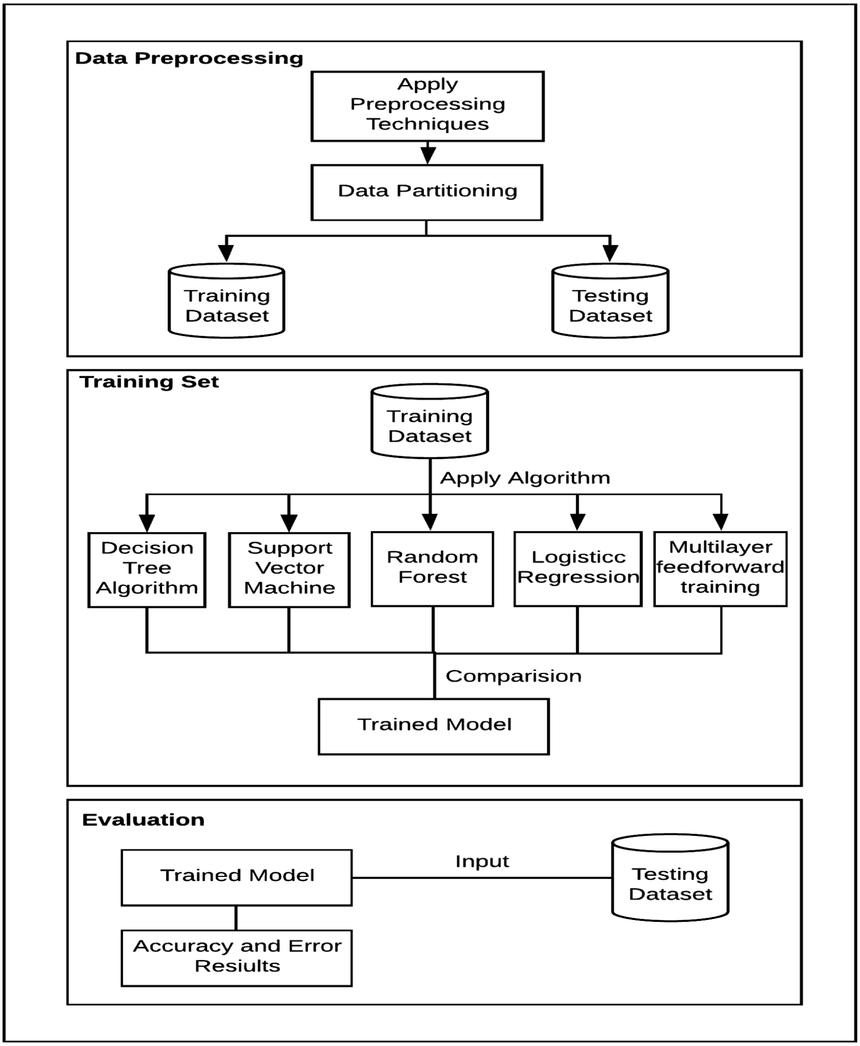

Fig. 1. System architecture

4.2 Process Diagram of Evaluation Method

Referring to the Fig. 2, initially, the data is collected from various sources. The collected data is then normalized. Then the dataset is split to training and testing dataset. For E.g. In decision tree method, prediction of class label for a record starts from the root of the tree. The values of the root attribute are compared with record's attribute. On the basis of comparison, the branch is followed to the corresponding value and jump to the next node. The comparison of the record's attribute with other internal nodes of the tree is continued until a leaf node occurs with predicted class value. The accuracy of each split is calculated using a function. The attribute with least cost is chosen for split. Using same strategy, the groups formed can be subdivided which makes this algorithm recursive in nature. As it has an excessive desire of lowering the cost, it is also called as the greedy algorithm. The root node is always the one which has the least gini score and thus the best classifier. Gini score gives the probability of choosing an item from the set and the probability of that item being misclassified.

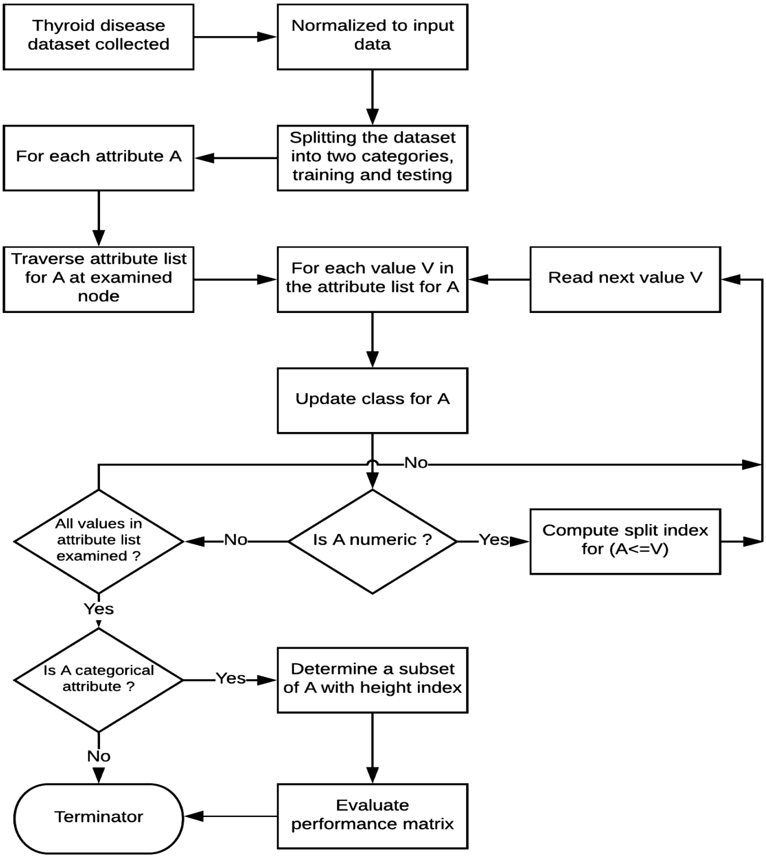
Thyroid disease dataset collected
Normalized to input data
Splitting the dataset into two categories, training and testing
For each attribute A
Traverse attribute list for A at examined node
For each value V in the attribute list for A
Read next value V
Update class for A
No
All values in attribute list examined ?
No
Is A numeric ?
Yes
Compute split index for (A<=V)
Yes
Is A categorical attribute ?
Yes
Determine a subset of A with height index
No
Terminator
Evaluate performance matrix

Fig. 2. Process diagram

Gini index is calculated by using:

$$Gini = 1 - \sum_j p_j^2$$

The value of gini index lies between 0 and 1. If the gini score is 0 then this indicates a perfect equality and if the gini score is 1, then this indicates a perfect inequality. The attribute with least gini index is chosen for the split.

4.3 Experimental Setup

Data Description. The dataset is obtained from several sources like thyroid disease dataset from UCI machine learning repository and other such repositories which consists of 10,450 records in total. There are total 29 features, out of which six features are real attributes and remaining are categorical attributes. To improve the quality of dataset obtained, pre-processing is carried out for further analysis (Table 1).

Table 1. Dataset description

Sr. no.	Attribute	Value type
1	Age	continuous
2	Sex	Male, Female
3	On thyroxine	False, True
4	query on thyroxine	False, True
5	on antithyroid medication	False, True
6	sick	False, True
7	pregnant	False, True
8	thyroid surgery	False, True
9	L131 treatment	False, True
10	query hypothyroid	False, True
11	query hyperthyroid	False, True
12	lithium	False, True
13	goitre	False, True
14	tumour	False, True
15	hypopituitary	False, True
16	psych	False, True
17	TSH measured	False, True
18	TSH	continuous
19	T3 measured	False, True
20	T3	continuous
21	TT4 measured	False, True
22	TT4	continuous
23	T4U measured	False, True
24	T4U	continuous
25	FTI measured	False, True
26	FTI	continuous
27	TBG measured	False, True
28	TBG	continuous
29	referral source	WEST, STMW, SVHC, SVI, SVHD, other

Implementation. The dataset collected from the source is classified using decision tree algorithm, random forest algorithm, support vector machine algorithm, logistic regression and multilayer feedforward algorithm. The accuracy for each algorithm is evaluated using performance matrix and the algorithm with highest accuracy is selected for future classifications.

5 Result and Analysis

The data obtained from the source was split in the ratio of 80:20. The training dataset consisted of 8360 training instances and the testing dataset consisted of 2090 testing instances. After training the model with all the 5 algorithms discussed above, the algorithms were tested using the testing dataset. The performance of the model was then evaluated via F1-score, Recall, Precision and Accuracy.

Figure 3 (bar graph to the left) below shows the comparison of the 5 algorithms discussed above on the basis of accuracy. Accuracy measures how well the system is performing under the testing data set. It is the ratio of accurately predicted samples to the total number of sample taken into consideration for the testing process.

$$\mathrm{Ac} = \mathrm{CP} + \mathrm{CN}/\mathrm{N}$$

Where,

Ac = accuracy
CP = correctly predicted positive samples
CN = correctly predicted negative samples
N = total Number of samples in the testing dataset

From the above comparison it is observed that Decision tree has the highest accuracy of 99.46%. Random forest algorithm has the accuracy very close to that of decision tree. The next algorithm logistic regression holds the accuracy of 97.5%. Support vector machine algorithm has the accuracy of 96.25% and multilayer feedforward obtains the last raking with the percentage equal to 95.17%.

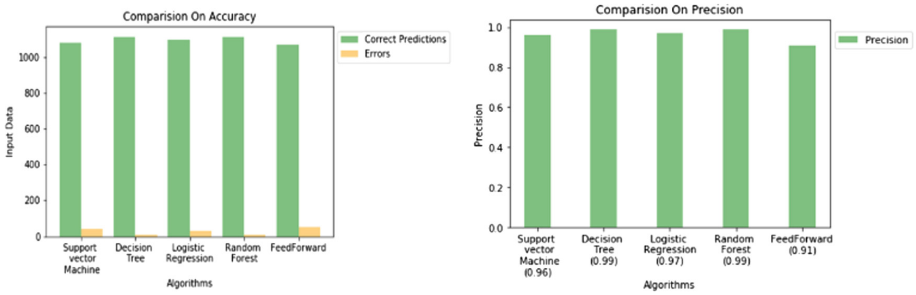

Fig. 3. Comparison on Accuracy and Precision

Figure 3 (bar graph to the right) shows the comparison on the basis of precision. Precision is calculated by considering the ratio of total number of positive samples to the total number of samples which the system reported to be positive

$$\mathrm{Pr} = \mathrm{P}/\mathrm{NP}$$

Where,

Pr = Precision
P = Positive samples
NP = total number of samples reported positive

Decision tree and the random forest algorithm got precision of 0.99 whereas logistic regression has precision of 0.97, support vector machine has precision of 0.96 and multilayer feedforward has least precision of 0.91.

Figure 4 (bar graph to the left) below shows the comparison on the basis of recall. Recall gives the measure of the system's positively predicted value to the total number of positive samples taken into consideration for testing process

$$\mathrm{Recall} = \mathrm{P}/\mathrm{NP}$$

Where,

P = truly predicted positive samples
NP = Total positive samples in the testing set

As it is observed from the Fig. 3, decision tree and random forest algorithm have a recall of 0.99, logistic regression has a recall of 0.97, support vector machine has a recall of 0.96, and multilayer feedforward has a recall of 0.95.

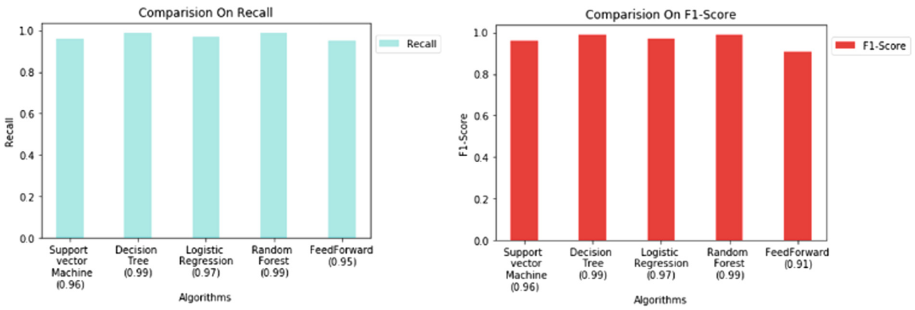

Fig. 4. Comparison on Recall and F1-score

Figure 4 (bar graph to the right) shows the comparison on the basis of F1-score. F1 scores measures the accuracy of the model by considering the Precision value and the recall values simultaneously.

$$\text{F1 score} = 2 * (\text{Re} * \text{Pr})/(\text{Re} + \text{Pr})$$

Where,

Re = Recall of the model
Pr = Precision of the model

It can be observed that decision tree and random forest have the highest F1-score of 0.99 whereas multilayer feedforward has the least F1-score of 0.91 (Table 2).

Table 2. Comparisons of various algorithms and parameters

	Accuracy	Precision	Recall	F1 score
SVM	96.25%	0.96	0.96	0.96
Decision Tree	99.46%	0.99	0.99	0.99
Logistic Regression	97.50%	0.97	0.97	0.97
Random Forest	99.30%	0.99	0.99	0.99
Feed Forward	95.17%	0.91	0.95	0.91

6 Conclusion

Thyroid is one of the most important gland of the body which controls metabolism and heart rate of our body. ‘T3’ and ‘T4’ are the hormones secreted by this gland which have a major role in controlling the metabolism and temperature of human body. In addition, 27 more parameters are considered in the proposed system using various machine learning algorithms in order to attain better accuracy in disease prediction. Also, after performing a comparative study to find the most accurate and precise algorithm for prediction, it can be concluded that decision tree algorithm gives the most accurate and precise results with the accuracy of 99.46% and precision of 0.99.

7 Future Scope

The proposed system may extended to the Internet of Things (IoT) concepts, which helps in real monitoring of thyroid patients and can predict diseases using ML techniques. Such interfaces provide greater help to patients as well as doctors and it will bring a revolutionary change in medical field as it can predict diseases with minimal errors and maximum efficiency. It makes system feasible for almost every type of user, including elderly and disabled persons.

References

1. Rao, N., Razia, S.: Machine learning techniques for thyroid disease diagnosis. Indian J. Sci. Technol. **9**(28) (2016). https://doi.org/10.17485/ijst/2016/v9i28/93705
2. Prerana, P.S., Taneja, K.: Predictive data mining for diagnosis of thyroid disease using neural network. Int. J. Res. Manag. Sci. Technol. **3**(2), 75–80 (2015). E-ISSN: 2321-3264
3. Umadevi, S., JeenMarseline, K.S.: Applying classification algorithms to predict thyroid disease. Int. J. Eng. Sci. Comput. **7**(10), 15118–15120 (2017)
4. Ammulu, K., Venugopal, T.: Thyroid data prediction using data classification algorithm. Int. J. Innov. Res. Sci. Technol. **4**(2), 208–212 (2017)
5. Ahmed, J., Abdul Rehman Soomrani, M.: TDTD: Thyroid Disease Type Diagnostics. In: 2016, International Conference on Intelligent Systems Engineering (ICISE) (2016). https://doi.org/10.1109/INTELSE.2016.7475160
6. Mahajan, P., Madhe, S.: Hypo and hyperthyroid disorder detection from thermal images using Bayesian classifier. In: 2014 International Conference on Advances in Communication and Computing Technologies (2014). https://doi.org/10.1109/EIC.2015.7230721
7. Saiti, F., Naini, A.A., Tehran, I., Shoorehdeli, M.A., Teshnehlab, M.: Thyroid disease diagnosis based on genetic algorithms using PNN and VM. In: 2009 3rd International Conference on Bioinformatics and Biomedical Engineering (2009). https://doi.org/10.1109/ICBBE.2009.5163689
8. Decision Trees in Machine Learning. https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052
9. Thyroid disorders in India: An epidemiological perspective. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3169866/#ref1
10. 7 Types of Artificial Neural Networks for Natural Language Processing. https://medium.com/@datamonsters/artificial-neural-networks-for-natural-language-processing-part-1-64ca9ebfa3b2
11. Thyroid Disease Data Set. http://archive.ics.uci.edu/ml/datasets/thyroid+disease