Thyroid Disease Prediction Using Machine Learning

Abstract

This paper is being written to provide a source of reference for the research scholars who want to work in the area of prediction of thyroid disease. From the different machine learning techniques, compared widely used three algorithms namely logistic regression, decision trees and k- nearest neighbor (kNN) algorithms to predict and evaluate their performance in terms of accuracy. This study has represented the intuition of how to predict the thyroid disease and highlighted how to apply the logistic regres- sion, decision trees and kNN as a tool for the classification. For this, thyroid data set of machine learning repository has used from UC Irvin knowledge discovery in databases archive.

Introduction

At least a person out of ten is suffered from thyroid disease in India. The disorder of thyroid disease primarily happens in the women having the age of 17–54. The extreme stage of thyroid results in cardiovascular complications, increase in blood pressure, maximizes the cholesterol level, depression and decreased fertility [1].

The hormones, total serum thyroxin (T4) and total serum triiodothyronine (T3) are the two active thyroid hormones produced by the thyroid gland to control the metabolism of body. For the functioning of each cell and each tissue and organ in a right way, in overall energy yield and regulation and to generate proteins in the ordnance of body temperature, these hormones are necessary [2, 3].

The idea for thyroid disease diagnosis and therapy is represented by the functional behavior of the thyroid dis- ease and is the key in most thyroid diseases. The basis of classification of thyroid disease is euthyroidism, hyper- thyroidism and hypothyroidism which are denoting normal, excessive or defective levels of thyroid hormones. The state euthyroidism depicts the normal production of thyroid hormones and normal levels at the cellular level by the thyroid gland. The state hyperthyroidism is clinical symptom due to excessive circulation and intracellular thyroid hormones. The state hypothyroidism is most of due to the lack of thyroid hormone generation and poor alter- nate therapy [4].

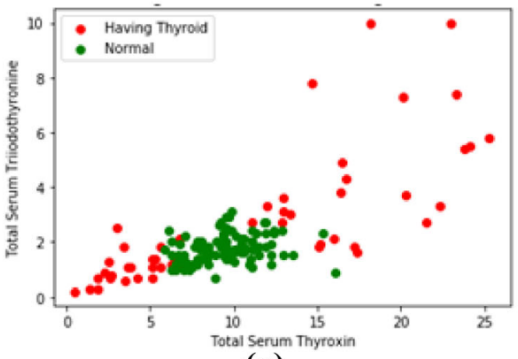
Cure of disease is a regular concern for the health care practitioners, and the errorless diagnostic at the right time for a patient is very important. Recently, by some advanced diagnosis methods, the common medical report can be generated with an additional report based on symptoms. The different questions like ‘‘what are the causes for affecting the thyroid?’’, ‘‘Which age group of people are affected due to thyroid?’’, ‘‘what is the relevant treatment for a disease?’’, etc. may find answers on implementing machine learning methods. Health care data can be pro- cessed and after implementing with certain methodologies; it can provide information that can be used in diagnosis and treatment of diseases more efficiently and accurately with better decision making and minimizing the death risk [5]. The large amount of data can be handled using the machine learning techniques.

Classification models are well suited for the classification and distinction of the data classes. The handling of both numerical and categorical values can be done by the classification processes. Classi- fication is a two-step classification model in the step one, based on some training data, a model is constructed, and in step two, an unknown tuple is given to the model to clas- sify into a class label [6].

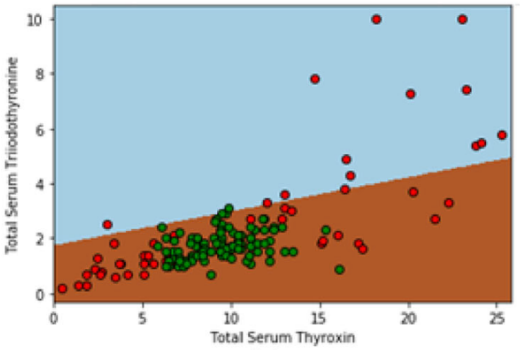
In human life, the classification has a great influence. The comparison of different classification techniques is a non-trivial and has a great dependency on the data set properties. In the statistics community, logistic regression, decision tree and k-nearest neighbor have got an esteemed position for classification problems [7].

Following steps are used to make a decision tree:

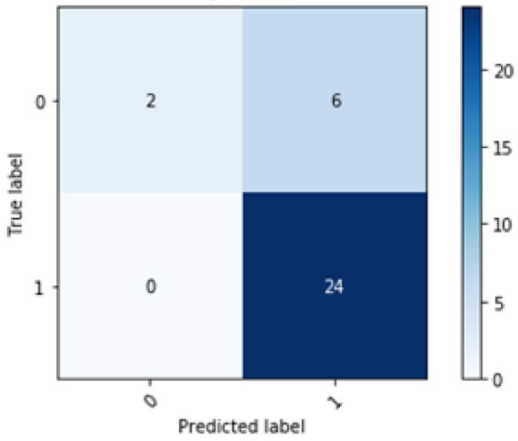
- Data preparation
- Data partition into training, validation and testing set
- Selection of attribute: a method to select the ‘‘best’’ possible attribute for the splitting by the decision tree model
- Evaluation of the model


Having Thyroid
Normal
Total Serum Triiodothyronine
Total Serum Thyroxin

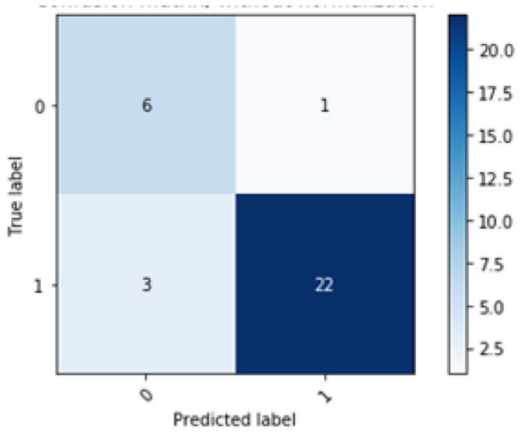
(a)


Total Serum Triiodothyronine
Total Serum Thyroxin

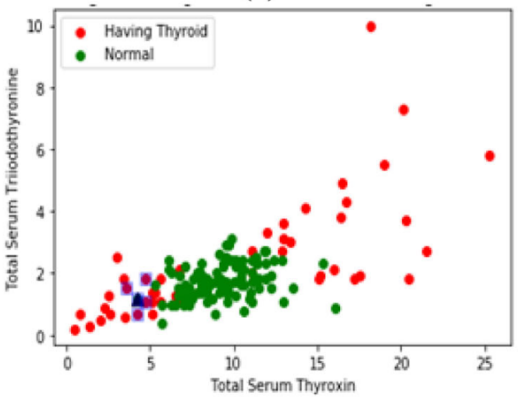
(b)


2
6
0
24
True label
Predicted label

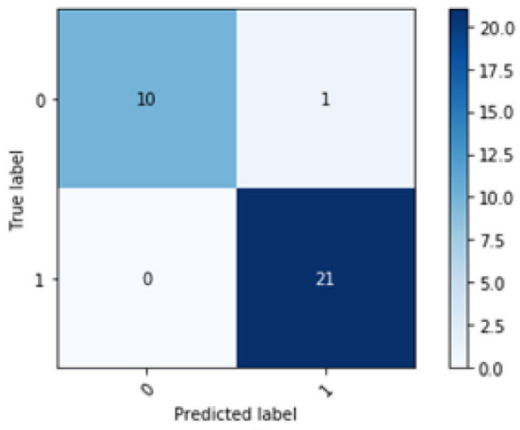
(c)


6
1
3
22
True label
Predicted label

(d)


Having Thyroid
Normal
Total Serum Triiodothyronine
Total Serum Thyroxin

(e)


10
1
0
21
True label
Predicted label

(f)