CORONARY HEART DISEASE PREDICTION

U15CS705R - COMPREHENSION AND TECHNICAL REPORT

Activity 2

submitted by

Sundhar U M (1517102163)

Sivanandham S (1517102151)

Tamilarasan (1517102166)

VII Semester

COMPUTER SCIENCE AND ENGINEERING

SONA COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

ANNA UNIVERSITY: CHENNAI 600 025

December 2020

ANNA UNIVERSITY: CHENNAI – 600 025

BONAFIDE CERTIFICATE

This is to certify that the technical report entitled **"Coronary Heart Disease Prediction"** is the bonafide report of **Sundhar U M (1517102163), S (15171021), Tamilarasan (1517102166)"** of B.E Computer Science and Engineering during the year 2020-21.

SIGNATURE
Dr.B.SATHIYABHAMA M.Tech,Ph.D.,
HEAD OF THE DEPARTMENT
Professor
Department of Computer Science and Engineering,
Sona College of Technology,
Salem.

SIGNATURE
Dr.S.SAKTHIVEL M.E,Ph.D.,
STAFF IN-CHARGE
Professor
Department of Computer Science and Engineering,
Sona College of Technology,
Salem.

Submitted for Comprehension and Technical Report **(U15CS705R)** examination held on (date of examination).

Examiner

CORONARY HEART DISEASE PREDICTION

Sundhar U M, Sivanandham S, Tamilarasan C
Final Year, Department of CSE,
Sona College of Technology,
Salem - 636005

ABSTRACT

Nowadays, health disease are increasing day by day due to life style, hereditary. Especially, heart disease has become more common these days, i.e. life of people is at risk. Each individual has different values for Blood pressure, cholesterol and pulse rate. But according to medically proven results the normal values of Blood pressure is 120/90, cholesterol is and pulse rate is 72. This paper gives the survey about different classification techniques used for predicting the risk level of each person based on age, gender, Blood pressure, cholesterol, pulse rate. The patient risk level is classified using datamining classification techniques such as Naïve Bayes, KNN, Decision Tree Algorithm, Random Forest. etc., Accuracy of the risk level is high when using more number of attributes.

INTRODUCTION

The Heart is one of the main organs of the human body. It pumps blood trough the blood vessels of the circulatory System. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Heart diseases or cardiovascular diseases (CVD) are a class of diseases that involve the heart and blood vessels. Cardiovascular disease includes coronary artery diseases (CAD) like angina and myocardial infraction. There is another heart disease, called coronary heart disease (CHD), in which a waxy substance called plaque develops inside the coronary arteries. These are the arteries which supply oxygen-rich blood to heart muscle. When plaque begins to build up in these arteries, the condition is called atherosclerosis. The development of plaque occurs over many years. With the passage of time, this plaque can harden or rupture (break open). Hardened plaque eventually narrows the coronary arteries which in turn reduces the flow of oxygen-rich blood to the heart. If this plaque ruptures, a blood clot can form on its surface. A large blood clot can most of the time completely block blood flow through a coronary artery. Over time, the ruptured plaque also hardens and narrows the coronary arteries. If the stopped blood flow isn't restored quickly, the section of heart muscle begins to die. Without quick treatment, a heart attack can lead to serious health problems and even death. We

are trying to predict the CHD using Machine Learning Models, Which will give us the result that the patient is prone to CHD or not.

DATA DESCRIPTION

The Dataset consists of 16 features,

I. Age – The age of the Patient.

II. Gender – '1' represents male and '0' represents female.

III. Current Smoker – '1' represents yes, '0' represents No.

IV. Prevalent Stroke – '1' represents that the patient had prevalent stroke, '0' represents that the patient had no strokes in past.

V. Prevalent HyperTension – '1' represents yes, '0' represents no.

VI. Diabetes – '1' represents yes, '0' represents no.

VII. Systolic Blood Pressure – The patients highest blood pressure value.

VIII. Diastolic Blood Pressure – The patients lowest blood pressure value.

IX. Heart Rate - Heart rate is the speed of the heartbeat measured by the number of contractions (beats) of the heart per minute value.

X. Education – '1' represents High school, '2' represents Diplamo, '3' represents College.

XI. Cigarettes Per Day – The Average count of cigarettes taken by the person.

XII. BP Medication – '1' represents that the patient is on BP Medication, '0' represents that the patient is not in BP Medication.

XIII. Total Cholesterol – The total cholesterol of the body.

BMI – The Body Mass Index (BMI) value of the person.

XIV. Glucose - Blood glucose level is the concentration of glucose level in the patient.

XV. Ten Year CHD – The patient record of having CHD disease. '1' represents yes, '0' represents no.

LITERATURE SURVEY

In [1] Shaikh Abdul Hannan et al. used a Radial Basis Function (RBF) to predict the medical prescription for heart disease. About 300 patient's data were collected from the Sahara Hospital, Aurangabad. RBFNN (Radial Basis Function–Neural Network) can be described as a three-layer feed forward structure. The three layers are the input layer, hidden layer and output layer. The hidden layer consists of a number of RBF units (*n*h) and bias (bk). Each neuron on the hidden layer uses a radial basis function as a nonlinear transfer function to operate on the input data. The most often used RBF is usually a Gaussian function. Designing a RBFNN involves selecting centres, number of hidden layer units, width and weights. The various ways of selecting the centres are random subset selection, k- means clustering and others. The

methodology was applied in MATLAB. Obtained results show that radial basis function can be successfully used (with an accuracy of 90 to 97%) for prescribing the medicines for heart disease**.**

In [2] Asha Rajkumar et al. [11] worked on diagnosis of heart disease using classification based on supervised machine learning. Tanagra tool is used to classify the data, 10 fold cross validation is used to evaluate the data and the results are compared. Tanagra is a free data mining software for academic and research purposes. It suggests several data mining methods from explanatory data analysis, statistical learning, machine learning and database area. The dataset is divided into two parts, 80% data is used for training and 20% for testing. Among the three techniques, Naïve Bayes shows lower error ratio and takes the least amount of time. Where accuracy were Naïve Bayes Scored 52.33%, Support Vector Machine scored 51%.

GAP ANALYSIS

In our proposed system we have applied different machine learning algorithms such as Naïve bayes, KNN, Decision tree Random Forest on a single dataset which is a bit larger than the dataset used by Shaikh Abdul Hannan. We predicted the accuracy of each model and compared to derive a better CHD model. And also we have done hyperparameter tuning using Random Search CV to improve the performance of the model.

ALGORITHMS

DECISION TREE - A Decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences including chance event outcomes and utility. It is one of the ways to display an algorithm. Decision trees are commonly used in operations research, specifically in decision analysis to help and identify a strategy that will most likely reach the goal. It is also a popular tool in machine learning. A Decision tree can easily be transformed to a set of rules by mapping from the root node to the leaf

nodes one by one. Finally, by following these rules, appropriate conclusions can be reached.

NAÏVE BAYES - It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers assume that the value of any particular feature is independent of the value of any other feature, given the class variable. Bayes theorem is given as follows: P(C|X) = P(X|C) * P(C)/P(X), where X is the data tuple and C is the class such that P(X) is constant for all classes. Though it assumes an unrealistic condition that attribute values are conditionally independent, it performs surprisingly well on large datasets where this condition is assumed and holds.

RANDOM FOREST - Random Forests are an ensemble learning method (also thought of as a form of nearest neighbour predictor) for classification and regression techniques. It constructs a number of Decision trees at training time and outputs the class that is the mode of the classes output by individual trees. It also tries to minimize the problems of high variance and high bias by averaging to find a natural balance between the two extremes. Both R and Python have robust packages to implement this algorithm.

K NEAREST NEIGHBOUR – A KNN Algorithm is an approach to data

classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in. The KNN is an example of a lazy learner algorithm, meaning that it does not build a model using the training set until a query of the data is performed.

PROPOSED SYSTEM

In our proposed system we have applied different machine learning algorithms such as Logistic regression, Naïve bayes, KNN, Random Forest, Decision tree etc on a single dataset and predicted the accuracy of each and compared to derive a better CHD model. Fig 1.1 shows the flow of the model building.

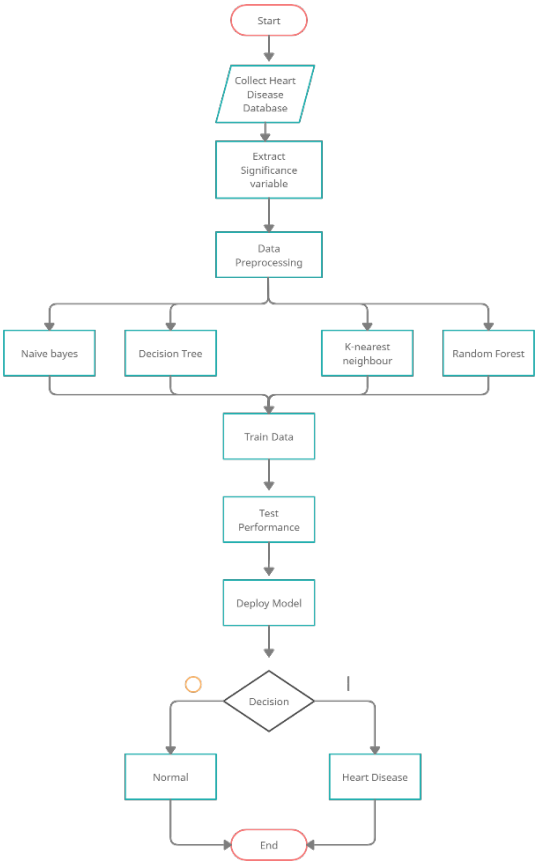

Fig-1.1 Flow Chart of model building

Steps involved in model building:

I. Import the required packages that is numpy, pandas, seaborn and matplotlib.pyplot.

II. A proper dataset must be collected and should ensure that there are no null values.

III. Import Dataset using pandas.

IV. Ensure thar there are no null values, if there is any fill them with random sample imputation method.

V. Separate the independent and dependent values and store them in ‘x’ and ‘y’ respectively.

VI. From sklearn import train_test_split method for further process.

VII. Split the dataset for training and testing using train_test_split method and give 80% of data for training the model and the rest can be used for testing.

VIII. Now choose the algorithms Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbour, Logistic regression form sklearn library, then train each model using 'fit' method and predict the test set separately for each algorithms.

IX. Evaluate the algorithms using confusion matrix and accuracy score and select the algorithm with higher accuracy.

X. After Selecting the best algorithm suitable, do hyper parameter tuning using 'Random Search CV' to improve the performance of the model.

ADVANTAGES

I. With this Analytics we can provide better information to the doctors.

II. Algorithms can provide immediate benefit to disciplines with processes that are reproducible or standardized.

III. Reducing healthcare costs, thus improving the quality of life.

IV. It is more efficient and faster than the conventional method which helps in reducing mortality rate.

V. This can be used in all Healthcare centre's which helps in diagnosing the CHD in a much efficient way to improve the health standards and provide required treatment thus reducing the mortality rate.

CONCLUSION

This paper presents a systematic review of different types of machine learning techniques to predict the heart disease. An analytical study has been conducted for the existing and accurate systems. On testing the dataset with different Algorithms it is founded that the Random Forest Classifier gives the maximum accuracy. The accuracy depends on the dataset provided.

REFERENCES

[1] Krish C Naik: Hyperparameter Tuning for an Algorithm, Random Sample Imputation.

[2] J Thomas & R Theresa Princy: Human Heart Disease Prediction using Data Mining Techniques.

[3] Rajesh Nichenametla, T.Maneesha, Shaik Hafeez, Hari Krishna: Prediction of Heart Disease Using Machine Learning Algorithms.

[4] Shaikh Abdul Hannan, Heart Disease Prediction using Artificial Neural Network.

[5] Animesh Hazra, Subrata Kumar & Amit Gupta, Heart Disease Diagonsis and Prediction using Machine Learning and Data Mining Techniques.