HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

[1] SHYLAJA.B , [2] LAKSHMI.C.N, [3] BINDHUSHREE.M,
[4] JAYA POOJARY ,[5] MANISH.C

[1] Asst. Professor, Department of Computer Science and Engineering
[2][3][4][5] BE Students, Department of Computer Science and Engineering
[1][2][3][4][5] Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India

Abstract

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyse huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as Naïve Bayes, decision tree, K-nearest neighbor, Support vector machine and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart disease patients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of diferent algorithms. This research paper aims to envision the probability of developing heart disease in the patients.

INTRODUCTION

Introduction Over the last decade, heart disease or cardiovascular remains the primary basis of death worldwide. An estimate by the World Health Organization, that over 17.9 million deaths occur every year worldwide because of cardiovascular disease, and of these deaths, 80% are because of coronary artery disease and cerebral stroke [1]. The vast number of deaths is common amongst low and middle-income countries [2]. Many predisposing factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various habitual risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are predisposing factors for heart disease. The efficient and accurate and early medical diagnosis of heart disease plays a crucial role in taking preventive measures to prevent death.

Data mining is exploring huge datasets to extract hidden crucial decisionmaking information from a collection of a past repository for future analysis. The medical field comprises tremendous data of patients. These data need mining by various machine learning algorithms. Healthcare professionals do analysis of these data to achieve effective diagnostic decision by healthcare professionals. Medical data mining using classification algorithms provides clinical aid through analysis. It tests the classification algorithms to predict heart disease in patients

Data mining is the process of extracting valuable data and information from huge databases. Various data mining techniques such as regression, clustering, association rule and classification techniques like Naïve Bayes, decision tree, random forest and K-nearest neighbor are used to classify various heart disease attributes in predicting heart disease. A comparative analysis of the classification techniques is used . In this research, I have taken dataset from the UCI repository. The classification model is developed using classification algorithms for prediction of heart disease. In this research, a discussion of algorithms used for heart disease prediction, comparison among the existing systems is made.

LITERATURE SURVEY:

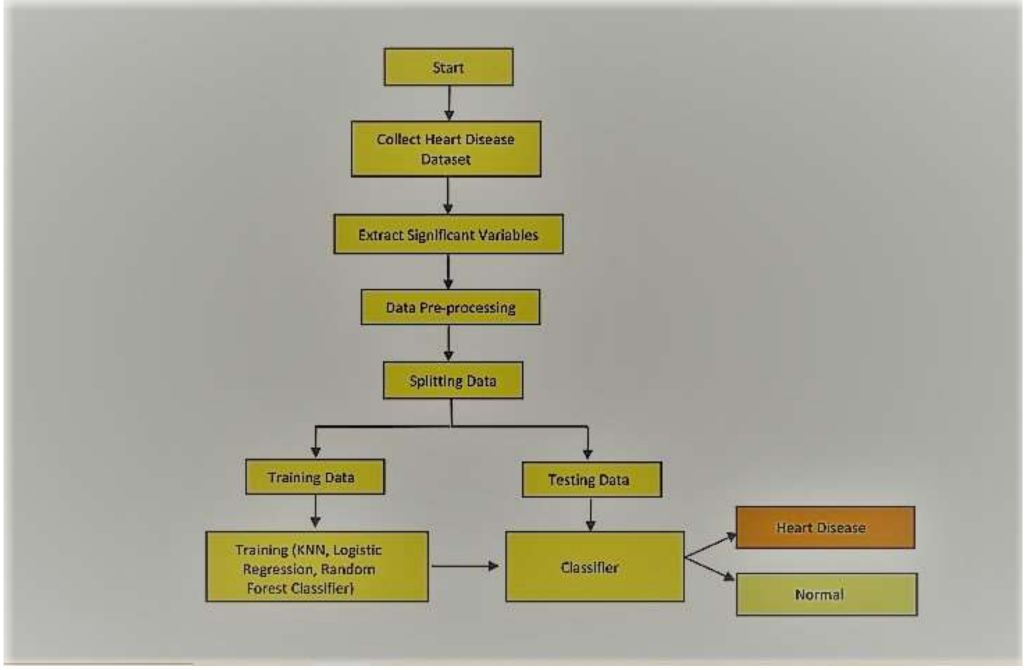
There is number of works has been done related to disease prediction systems using different machine learning algorithmsin medical Centre's.

Author	Year of publication	Title	Objective	Dataset used	Result
Rohit Bharti, Aditya Khamparia, Mohammad Shabaz Gaurav Dhiman	Jul 2021	Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning	Different machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset.	UCI Machine Learning Heart Disease dataset . Consisting of 14 attributes.	Random Forest 76.7%, Logistic Regression 83.64%, KNNeighbors 82.27%, Support Vector Machine 84.09%, Decision Tree 75.0%, and XGBoost is 70.0%.
Harshit Jindal, Sarthak Agrawal, Rishabh Khera	Jan 2021	Heart disease prediction using machine learning algorithms.	Prepared heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient.	Dataset is selected from UCI repository.	Maximum accuracy obtained from KNN and logistic regression is 88.5.
Dhai Eddine Salhi, Abdelkamel Tari, and M-Tahar Kechadi3	Feb 2021	Using machine learning for heart disease prediction.	Used data analytics to detect and predict disease's patients. Starting with a preprocessing phase, where the most relevant features were selected then three data analytics techniques were applied on data sets of different sizes.	Structured data set of Algerian people.	Neural-networks (600 lines) 91.8% Neural-networks (800 lines) 92% Knn (600 lines) 85.1% Knn (800 lines) 85.3% Svm (600 lines) 89.7% Svm(800 lines) 89%

Baban Uttamrao Rindhe, Nikita Ahire, Rupali Patil	May 2021	Heart Disease Prediction Using Machine Learning	The main objective of this research project is to predict the heart disease of a patient using machine learning algorithms	UCI Machine learning repository dataset.	Support vector classifier :84% Neural network 83.5% Random forest classifier :80%
Mangesh limbiote, Kedhar damkondavar, Pushkar patil	June 2020	Survey on prediction techniques of heart disease using machine learning.	In-depth analysis of relevant ml techniques to predict heart disease.	UCI machine learning repository	Svm:82.30% Random forest:91.3%

PROPOSED METHODOLOGY

Processed methodology start with the collection of data for this download tha data from kaggle that is well verified by researchers. In This methodology, There are many steps as shown in block diagram.


Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Splitting Data
Training Data
Testing Data
Training (KNN, Logistic Regression, Random Forest Classifier)
Classifier
Heart Disease
Normal

Description of the Dataset:

The dataset used for this research purpose was the Public Health Dataset and it is dating from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The “target” field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 =disease. The first four rows and all the dataset features are shown in Table 1 without any preprocessing. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:

(i)Age—age of patient in years, sex—(1 =male; 0 =female).

(ii) Cp—chest pain type.

(iii) Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little

(iv) higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).

(v) Chol—serum cholesterol shows the amount of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).

(vi) Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.

(vii) Restecg—resting electrocardiographic results.

(vii)Thalach—maximum heart rate achieved.The maximum heart rate is 220 minus your age.

(viii) Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.

(ix)Oldpeak—ST depression induced by exercise relative to rest.

(x) Slope—the slope of the peak exercise ST segment.

(xi) Ca—number of major vessels (0–3) colored by fluoroscopy

(xii) Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).

(xiii) Target (T)—no disease = 0 and disease =1, (angiographic disease status).

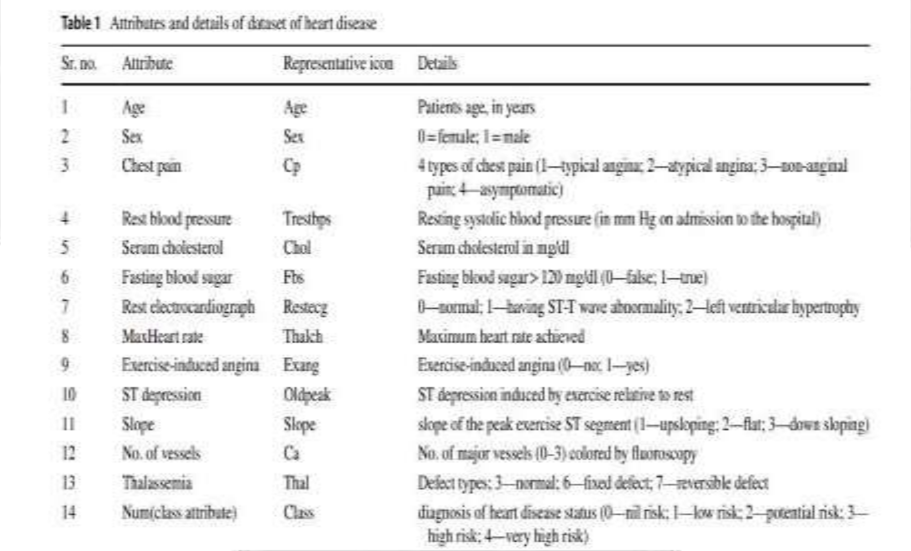


Table 1 Attributes and details of dataset of heart disease

Sr. no.	Attribute	Representative icon	Details
1	Age	Age	Patients age, in years
2	Sex	Sex	0 = female; 1 = male
3	Chest pain	Cp	4 types of chest pain (1—typical angina; 2—atypical angina; 3—non-anginal pain; 4—asymptomatic)
4	Rest blood pressure	Trestbps	Resting systolic blood pressure (in mm Hg on admission to the hospital)
5	Serum cholesterol	Chol	Serum cholesterol in mg/dl
6	Fasting blood sugar	Fbs	Fasting blood sugar > 120 mg/dl (0—false; 1—true)
7	Rest electrocardiograph	Restecg	0—normal; 1—having ST-T wave abnormality; 2—left ventricular hypertrophy
8	MaxHeart rate	Thalch	Maximum heart rate achieved
9	Exercise-induced angina	Exang	Exercise-induced angina (0—no; 1—yes)
10	ST depression	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope	slope of the peak exercise ST segment (1—upsloping; 2—flat; 3—down sloping)
12	No. of vessels	Ca	No. of major vessels (0–3) colored by fluoroscopy
13	Thalassemia	Thal	Defect types; 3—normal; 6—fixed defect; 7—reversible defect
14	Num(class attribute)	Class	diagnosis of heart disease status (0—nil risk; 1—low risk; 2—potential risk; 3—high risk; 4—very high risk)

Data Pre-processing:

The real-life information contains large numbers with missing and noisy data. These data are pre-processed to overcome such issues and make predictions vigorously. Figure 1 explains the sequential chart of our proposed model. Cleaning the collected data usually has noise and missing values. To get an accurate and effective result, these data need to be cleaned in terms of noise and missing values are to be filled up. Transformation it changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks.

Integration the data may not be acquired from a single source but varied sources, and it has to be integrated before processing. Reduction the data gained are complex and require to be formatted to achieve efective results. The data are then classifed and split into training data set and test data set which is run on various algorithms to achieve accuracy score results.

MACHINE LEARNING ALGORITHMS

Naïve bayes[NB] :NB is a supervise classification algorithm. It is a simple technique using Bayes theorem. To get the probability, mathematical concept is used with the support of bayes theorem. The correlation is neither related to each other nor predictor to one another. All parameters work autonomously for getting the maximum probability. $P(x/y) = P(Y/X)\ P(x) / p(y)$ Where p(x)=Class predictor probability, p(y)= Predictor Probability,

P(x/y)= Posterior probability,

P(y/x)=possibility, probability of predictor

Decision Tree[DT] :

DT is an algorithm that classifies parameters in categorical form in spite of arithmetic data. Tree like structure is created by DT. Many large data set related to medical have analyzed by DT due to its simple nature. It works on tree node for analysis. Leaf Node: Signify the solution of every Test Interior Node: Handle numerous element Main Node[Root Node]: Other nodes work based on main node Data is to be divided into two or more parallel set by applying this algorithm. Then entropy of each parameter is calculated. After that divide the data with predictor having extreme information gain that means minimum entropy

Entropy(S) = $\sum$c i=1 −Pi log2 Pi,

Gain (S, A) = Entropy(S) − $\sum$ v Values(A) |Sv| |S| Entropy (Sv).

3. Random forest [RF] :

RF algorithm is supervised primarily based learning. It is used as classifier in numerous fields. By using this more trees makes a forest. If we have more number of trees then it create higher accuracy. It is also used for regression task. but it accomplish well when classify the task. And may overwhelmed misplaced values. There are three approach of RF: Forest RC(Random Blend) Forest RI(Random input) And combination of RC and RI.

Logistic regression [LR]:

LR is the supervised ML learning method. It is established on the association between dependent and independent variable as seen in Fig.5 variable “a” and “b” are dependent variable and independent variable and relation between them is shown by equation of line which is linear in nature that why this approach is called linear regression.

4.Support Vector Machine :

SVM is one type of ML method that work on the conception of hyper plan. It is used to find a hyper plan in n dimensional space, using this data point can be classified specifically [13]. (Xa, Ya) is training sample of data set where a=1,2,3,…….n and Ya is the target vector and Xa is the ith vector. Hyper plan quantity select the variety of support vector such as example if a line is used as hyper plan then method is called linear support vector.

5.K-nearest Neighbor [KNN]:

KNN is a classification algorithm that belongs to supervise learning. It categorizes the entity that reliant on nearest neighbor. KNN could be a wide applied methodology used as a classifier and regression in numerous field like image process, data processing, pattern recognition and different applications. The output result of the algorithmic program depends on Knearest neighbor class that enforced by finding K- variety of coaching points nearest to the specified character and contemplate the votes among the K object. The algorithmic program is incredibly easy. However, is capable of learning highly-complex non-linear call boundaries and regression functions . The intuition of KNN that similar instances ought to have similar category labels (in classification) or similar target values (regression). On the drawback, the algorithmic program is computationally high-priced, and is vulnerable to over fitting.

RESULT AND ANALYSIS:

Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on super vised machine learning classification techniques using Naïve Bayes, decision tree, random forest, and K-nearest neighbor on UCI repository. Various experiments using different classifier algorithms were conducted through the WEKA tool. Research was performed on 8th generation Intel Corei7 having an 8750H processor up to 4.1 GHz CPU and 16 GB ram. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest are applied to get accuracy score. The accuracy score results of different classification techniques were noted using Python Programming for training and test data sets.

By applying different machine learning algorithms and then using deep learning to see what difference comes when it is applied to the data, three approaches were used. In the first approach, normal dataset which is acquired is directly used for classification, and in the second approach, the data with feature selection are taken care of and there is no outliers detection. The results which are achieved are quite promising and then in the third approach the dataset was normalized taking care of the outliers and feature selection; the results achieved are much better than the previous techniques, and when compared with other research accuracies, our results are quite promising

CONCLUSION:

Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on super vised machine learning classification techniques using Naïve Bayes, decision tree, random forest, support vector machine and K-nearest neighbor on UCI repository. Various experiments using diferent classifier algorithms were conducted through the WEKA tool. Research was performed on 8th generation Intel Corei7 having an 8750H processor up to 4.1 GHz CPU and 16 GB ram. Dataset was classified and split into a training set and a test set. Pre-processing of the data is done and supervised classification techniques such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest are applied to get accuracy score. The accuracy score results of diferent classification techniques were noted using Python Programming for training and test data sets.

REFERENCES:

Extensive study about the topic was performed and various methodologies used in this domain were found. Predominantly there were two types of analysis: exploratory and predictive. Exploratory analysis visualizes events that have occurred in the past and provides meaningful insights that can be used for decision making.

1. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3884968
2. https://www.hindawi.com/journals/cin/2021/8387680/
3. https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/meta
4. https://www.researchgate.net/publication/348604625_Heart_disease_prediction_using_machine_learning_algorith ms
5. https://www.researchgate.net/publication/349470771_Using_Machine_Learning_for_Heart_Disease_Prediction
6. https://deliverypdf.ssrn.com/delivery.php?ID=1090640311011230640210130730651060760990570860000170350670890991031130220911040021000560570250021101210521240190980850821120180220730380440320060980250680790730080660840770541040220710190920680201041000250071230980821210920980300871160080780750720121 09&EXT=pdf&INDEX=TRUE