https://doi.org/10.1007/s42979-020-00365-y

ORIGINAL RESEARCH

Heart Disease Prediction using Machine Learning Techniques

Parthiv Rawat[1]

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, 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 different algorithms. This research paper aims to envision the probability of developing heart disease in the patients. The results portray that the highest accuracy score is achieved with K-nearest neighbor.

Keywords Heart disease prediction · Data mining · Decision tree · Naïve Bayes · K-NN · Random forest · Machine learning

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 refers to the extraction of required information from huge datasets in various fields such as the medical field, business field, and educational field. Machine learning is one of the most rapidly evolving domains of artificial intelligence. These algorithms can analyse huge data from various fields, one such important field is the medical field. It is a substitute to routine prediction modeling approach using a computer to gain an understanding of complex and non-linear interactions among different factors by reducing the errors in predicted and factual outcomes [3]. Data mining is exploring huge datasets to extract hidden crucial decision-making 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

This article is part of the Machine Learning Course Project.

✉ Parthiv Rawat
2018176@iiitdmj.ac.in

[1] Mechanical Engineering Department, Pandit Dwarka Prashad Mishra Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005, India

classification algorithms to predict heart disease in patients [4].

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 [5]. 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. It also mentions further research and advancement possibilities in the paper.

Background

Heart disease affects millions of people, and it remains the chief cause of death in the world. Medical diagnosis should be proficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technology that helps computers to build and classify various attributes. This research paper uses classification techniques to predict heart disease. This section gives a portrayal of the related subjects like machine learning and its methods with brief descriptions, data pre-processing, evaluation measurements and description of the dataset used in this research.

Machine Learning

Machine learning is an emerging subdivision of artificial intelligence. Its primary focus is to design systems, allow them to learn and make predictions based on the experience. It trains machine learning algorithms using a training dataset to create a model. The model uses the new input data to predict heart disease. Using machine learning, it detects hidden patterns in the input dataset to build models. It makes accurate predictions for new datasets. The dataset is cleaned and missing values are filled. The model uses the new input data to predict heart disease and then tested for accuracy. Machine learning techniques are classified as:

Supervised Learning

The model is trained on a dataset that is labelled. It has input data and its outcomes. Data are classified and split into training and test dataset. Training dataset trains our model while testing dataset functions as new data to get accuracy of the model. The dataset exists with models and its output. The classification and regression are its example.

Unsupervised Learning

Data used to train are not classified or labelled in the dataset. Aim is to find hidden patterns in the data. The model is trained to develop patterns. It can easily predict hidden patterns for any new input dataset, but upon exploring data, it draws conclusion from datasets to describe hidden patterns. In this technique, no responses in the dataset are seen. The clustering method is an example of an unsupervised learning technique.

Reinforcement Learning

It does not use labelled dataset nor the results are associated with data, thus model learns from the experience. In this technique, the model improves its presentation based on its association with environment and figures out how to discuss its faults and to get the right outcome through assessment and testing various prospects.

Classification algorithms are commonly used supervised learning techniques to define probability of heart disease occurrence.

Classification Machine Learning Techniques

The classification task is used for prediction of subsequent cases dependent on past information. Many data mining techniques as Naïve Bayes, neural network, decision tree have been applied by researchers to have a precision diagnosis in heart disease. The accuracy given by different techniques varies with number of attributes. This research provides diagnostic accuracy score for improvement of better health results. We have used WEKA tool in this research for pre-processing the dataset, which is in ARFF format (attribute-relation file format). Only 14 attributes of all 76 different attributes have been considered for analysis to get precise results. By comparison and analysis using different algorithms with WEKA tool heart disease can be predicted and treated early and prompt [5].

Approach Methodology

This research aims to foresee the odds of having heart disease as probable cause of computerized prediction of heart disease that is helpful in the medical field for clinicians and patients [5]. To accomplish the aim, we have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in this research paper. This paper additionally depicts which attributes contribute more than the others to anticipation of higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome [5].

Data Source

For this study, I have used dataset from Kaggle Machine learning repository. It comprises a real dataset of 304 examples of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. (Table 1). In this research, we have used four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.

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, thes 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.

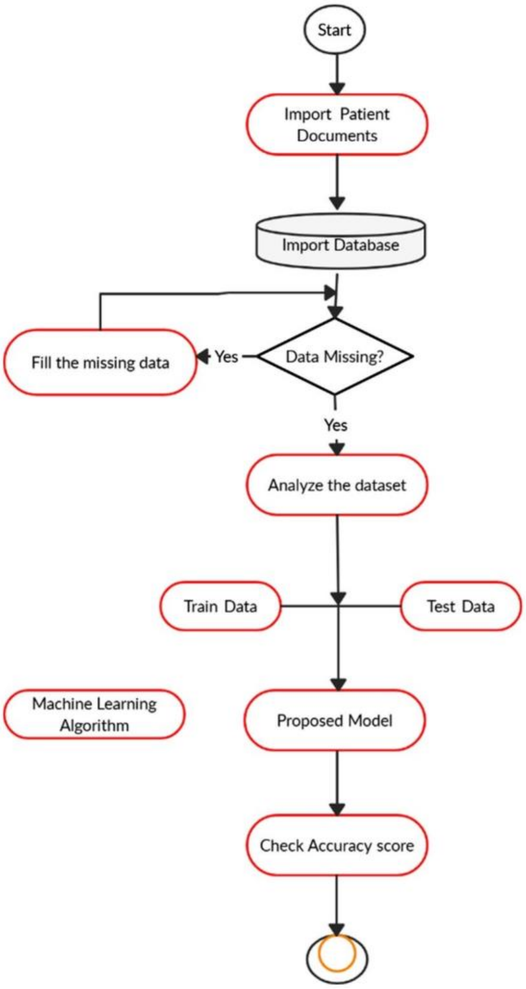


Fig. 1 Sequential chart of proposed model

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)

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 effective results.

The data are then classified and split into training data set and test data set which is run on various algorithms to achieve accuracy score results.

Algorithms Used

Naïve Bayes

Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes strong (Naive) independence among attributes. Bayes theorem is a mathematical concept to get the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. It is able to work with Naïve Bayes model and does not use Bayesian methods. Many complex real-world situations use Naive Bayes classifiers [6]:

$$P(X/Y) = \frac{P(Y/X) \times P(X)}{P(Y)},$$

$P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

Naïve Bayes is a simple, easy to implement, and efficient classification algorithm that handles non-linear, complicated data. However, there is a loss of accuracy as it is based on assumption and class conditional independence.

The accuracy of 85.25% has been achievedusing all 13 attributes of Cleveland dataset [8].

Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures. Decision tree is simple and widely used to handle medical dataset. It is easy to implement and analyse the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

- Root node: main node, based on this all other nodes functions.
- Interior node: handles various attributes.
- Leaf node: represent the result of each test.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

$$\text{Entropy}(S) = \sum_{i=1}^{c} -Pi \log_2 Pi,$$

$$\text{Gain }(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|Sv|}{|S|} \text{Entropy }(Sv).$$

The results obtained are easier to read and interpret [3]. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph.

An accuracy of 81.97% has been achieved by the decision tree. [9].

K-Nearest Neighbor (K-NN)

The K-nearest neighbors algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbor. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbors is measured using Euclidean distance [3]. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy. In a study 67.21% accuracy was achieved with value $K=9$ [8].

Random Forest Algorithm

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more number of trees give higher accuracy. The three common methodologies

- Forest RI (random input choice);
- Forest RC (random blend);
- Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Random forest algorithm has obtained an accuracy of 90.16% with Cleveland dataset.

Logistic Regression

It is a statistical model. A logistic regression is a classification algorithm. For binary classification problem, in order to predict the value of predictive variable y when $y \in [0, 1]$, 0 is negative class and 1 is positive class. It also uses multi classification to predict the value of y when $y \in [0, 1, 2, 3]$. In order to classify two classes 0 and 1, a hypothesis $h(\theta) = \theta\ TX\ X$ will be designed and threshold classifier output is $h\theta(x)$ at 0.5. If the value of hypothesis $h\theta\ (x) \geq 0.5$, it will predict y =1 which mean that the person has heart disease and if value of $h\theta(x) < 0.5$, then predict y = 0 which shows that the person is healthy. Hence, the prediction of logistic regression under the condition $0 \leq h\theta(x) \leq 1$ is done.

Logistic regression sigmoid function can be written as follows:

$$h\theta(x) = g\ \theta T - X\ ,$$

where $g(z)\ 1/(1 + x{-}z)$ and $h\theta(x) =1/(1 + x{-}z)$.

Similarly, the logistic regression cost function can be written as follows:

$$J(\theta) =1/m \sum m\ i{=}1\ cost(\ h\theta\ (\ x(i)\ )y(i)\ )$$

Logistic Regression algorithm has obtained an accuracy score of 85.25% with Cleveland dataset.

Support Vector Machine (SVM)

SVM is a machine learning algorithm used for classification/ regression .It classifies both linear and non-linear data. It separate data based on labels. The technique, kernel trick used to match new data to best from training data to predict unknown target label. Learn from past labeled data and predict future. Data from two classes can always be separated by a hyper plane. The SVM find this hyper plane using support vectors and margins. SVM performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors.

Binary classification problem, the instances are separated with a hyper plane $wTx+b = 0$, where w and dare dimensional coefficient vectors, which are normal to the hyperplane of the surface, b is off set value from the origin, and x is data set values. The SVM gets results of w and b. w can be solved by introducing Lagrangian multipliers in the linear case. The data points on borders are called support vectors .The solution of w can be written as :

$$w =\sum n\ i{=}1\ \alpha i\ yi\ xi$$

, where n is the number of support vectors and yi are target labels to x. The value of w and b are calculated, and the linear discriminant function can be written as follows:

$$g(x)= sgn(\sum n\ i\ {=}1\ \alpha i\ yi\ xi\ T\ x{+}b)$$

The nonlinear scenario, for kernel trick and decision function, can be written as follows:

$$g(x)= sgn(\sum n\ i{=}1\ \alpha i\ yi\ K(\ xi{+}\ x){+}b)$$

The SVM algorithm has obtained accuracy of 81.97% using the Cleveland data.

XGBoost

XGBoost is an open-sourced software which provides a gradient boosting framework for C++, Java, Python, R, Julia, Perl, and Scala. It works on Linux, Windows, and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, and GBDT) Library. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made. A popular example is the AdaBoost algorithm that weights data points that are hard to predict.

Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models. This approach supports both regression and classification predictive modeling problems.

The XGBoost software obtained a substantial accuracy of 78.9%.

Neural Network

A **neural network** is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes.[1] Thus a neural network is either a biological neural network, made up of real biological neurons, or an artificial neural network, for solving artificial intelligence (AI) problems. The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

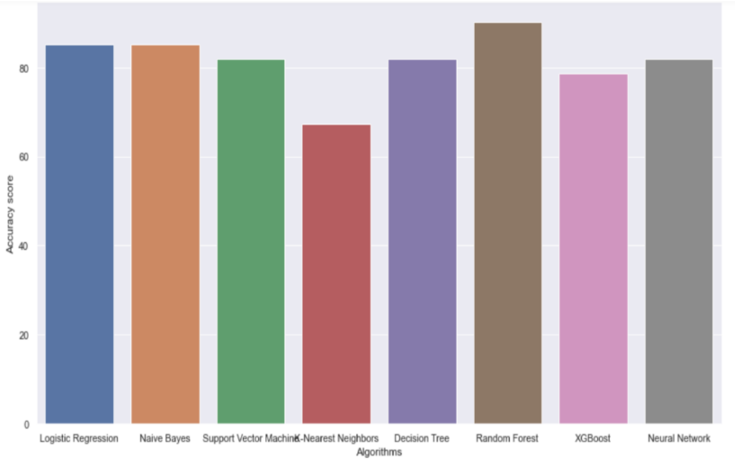
Machine Learning Algorithm	Logistic Regression	Decision Tree	K-Nearest Neighbor	Random Forest	Naïve Bayes	Support Vector Machine	XGBoost	Neural Network
Accuracy	85.25%	81.97%	67.21%	90.16%	85.25%	81.97%	78.69%	81.97%

Results and Analysis

Aim of this research is to predict whether or not a patient will develop heart disease. This research was done on supervised machine learning classification techniques using Naïve Bayes, decision tree, random forest, K-nearest neighbor, logistic regression, SVM, XGBoost and Neural Network 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 Programmingfor training and test data sets

Conclusion

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, I consider only 14 essential attributes. I applied four data mining classification techniques, K-nearest neighbor, Naive Bayes, decision tree, and random forest. The data were pre-processed and then used in the model. Logistic Regression, Naïve Bayes, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing eight algorithms to be highest in Random Forest. We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules, and genetic algorithm. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.



References

1. Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.
2. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low-and middle-income countries. Curr Probl Cardiol. 2010;35(2):72–115.
3. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.
4. Ramalingam VV, Dandapath A, Raja MK. Heart disease predic- tion using machine learning techniques: a survey. Int J Eng Tech- nol. 2018;7(2.8):684–7.
5. Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.
6. Fatima M, Pasha M. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9:1–16. https://doi.org/10.4236/jilsa.2017.91001.
7. Pahwa K, Kumar R. Prediction of heart disease using hybrid technique for selecting features. In: 2017 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON). IEEE. p. 500–504.
8. Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutier-rez J. A comprehensive investigation and comparison of machinelearning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p.204–207.
9. Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to pre-dict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.
10. Bouali H, Akaichi J. Comparative study of different classifica- tion techniques: heart disease use case. In: 2014 13th interna- tional conference on machine learning and applications. IEEE. p.482–86.