International Journal for Modern Trends in Science and Technology, 6(12): 290-293, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061254
Available online at: http://www.ijmtst.com/vol6issue12.html

Heart Disease Prediction Using Machine Learning

Tarun Rahuja [1] | Nidhi Sengar[2] | Dr.Amita Goel[3]

[1]B.Tech Scholar, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi, India
[2]Assistant Professor, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi, India
[3]Associate Professor, Department of IT, Maharaja Agrasen Institute Of Technology, Delhi, India

To Cite this Article
Tarun Rahuja, Nidhi Sengar and Dr.Amita Goel, "Heart Disease Prediction Using Machine Learning", *International Journal for Modern Trends in Science and Technology*, 6(12): 290-293, 2020.

Article Info
Received on 10-November-2020, Revised on 02-December-2020, Accepted on 06-December-2020, Published on 11-December-2020.

ABSTRACT

This paper revolves around a classification use case of machine learning in which the intention is to predict the possibility of a heart disease in an individual given certain parameters. Machine Learning is extensively being used across the world. The healthcare industry has also commenced leveraging these data driven techniques. Machine Learning can play a vital role in predicting the likelihood of locomotor disorders, Heart ailments and more such diseases because machine learning is well known for its use cases in classifying, categorizing and predicting. Such information, if predicted well, can provide key foresight to doctors who can hence mould their diagnosis and course of treatment per patient basis. The main advantage of using machine learning in healthcare is its ability to parse and process huge datasets which are beyond the scope of human abilities, and then accurately convert the derived analysis of that data into clinical insights that can aid medical practitioners round the globe in planning stratergies for providing care to patients, ultimately leading to more promising results, reduced costs of care and last but not the least , increased patient satiation and response/ recovery. To simplify and solve this problem, solutions were provided using multiple supervised learning algorithms like logistic regression, Naïve Bayes, random forests, decision trees, support vector machines and K-nearest neighbours. The best accuracy was seen using random forests.

Keywords : *Machine learning, supervised learning, logistic regression, Naïve Bayes, random forests, decision trees, support vector machines, K-nearest neighbours*

INTRODUCTION

Heart is an organ of paramount importance in the human body. Its core function is to regulate and manipulate the blood flow to the various parts of the human body. In case of an ailment to the heart this function of the heart could get affected which can lead to life threatening conditions like stroke and paralysis. In today's fast paced environment a large part of our population is exposed to situations of extreme stress and hostile work environments which has led to a significant increase in the instances of heart ailments. Heart disease in most cases is triggered due to unhealthy lifestyle, smoking, intake of alcohol and excessive intake of fat which may cause plaque build up and narrowing of blood vessels triggering hypertension. According to the World Health Organization more than 10 million die due to Heart diseases every single year around the world. Heart diseases have emerged as one of the most prominent cause of death all around the world. According to the organisation, heart related diseases are responsible for 31% of all global deaths. In India too, heart related diseases have become the leading cause of mortality. Estimates made by the World Health Organisation (WHO), suggest that India have lost up to $237 billion, from 2005-2015, due to heart related or Cardiovascular diseases [2]. Thus, viable and precise prediction of heart related diseases is of

paramount importance. A healthy lifestyle and early stage detection are the only ways to prevent the heart related diseases. Medical institutions all around the world follow a practice of accumulating data about patients and medical ailments so that the collected data can be used to derive clinical insights but unfortunately this data is mostly unstructured and tedious to be interpreted by the human mind since it is very large in amount and most of the times comprises of noise and anomalies that need to cleaned beforehand. This is where machine learning comes into picture, we can use number crunching algorithms on this data to perform meaningful computations and derive key insights and conclusions . Data mining is done to extract insights out of this dataset. Data mining is the process of extracting valuable 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 medical attributes about the subject in predicting heart disease. A comparative analysis of the classification techniques is used [3]. This paper makes use of heart disease dataset available in UCI machine learning repository.

II. RELATED WORKS

There have been many researches carried out in the field of integrating medical science and machine learning. There have been multiple studies related to disease prediction using machine learning. T. R. Reed, N. E. Reed, and P. Fritzson, in their paper, Heart sound analysis for symptom detection and computer-aided diagnosis described a technique to generate artificial sounds mimicking those that are produced by the heart to allow practitioners aquire the skill of heart auscultation to accurately predict those diseases that cannot be traced via modern techniques like electrocardiography [4]. S. Mohan, Chandrasegar Thirumalai, G. Srivastava in their paper proposed a novel method that aimed to find substantial features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease [5]. Lamido Yahaya , Nathaniel David Oye, Etemi Joshua Garba in their paper, A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques, investigated the state of various clinical decision systems for heart disease prediction, proposed by various researchers using data mining and machine learning techniques. Classification algorithms such as the Naïve Bayes (NB), Decision Tree (DT), and Artificial Neural Network (ANN) were employed to predict heart diseases to obtain varying degrees of accuracy [6].

III. METHODOLOGY

Dataset acquisition : The dataset used was the Heart disease Dataset which is a combination of 4 different databases, but only the UCI Cleveland dataset has been used. This database consists of a total of 76 attributes but all published experiments use only a subset of only 14 features [7]. Out of these 14 attributes there are 13 predictor variables and 1 class label. The various features of the dataset have been described below :

Table 1. Attribute Description of Dataset

S.No	Attribute Name	Description	Domain
1)	Age	Current age of subject	Numerical
2)	Sex	Gender	Binary
3)	Chp	Chest pain type	Categorical
4)	Bp	Resting Blood pressure	Numerical
5)	SCh	Serum Cholestrol	Numerical
6)	ECG	Resting electrocardiograph result	Categorical
7)	Mhrt	Maximum Heart Rate	Numerical
8)	Exian	Exercise induced angina	Binary
9)	FBS	Fasting Blood Sugar > 120 mg/dl	Binary
10)	Opk	Old peak	Numerical
11)	Slope	Slope of peak exercise ST segment	Categorical
12)	Vessel	Number of vessels coloured by fluoroscopy	Categorical
13)	Thal	Defect type	Nominal
14)	Class Label	Presence/Absence of heart disease	Binary

Data Preprocessing : Data that we want to process might not be in the form that is immediately usable that is it may contain noise or it may have missing values, in such a scenario, to obtain accurate

results we need to eliminate these anomalies in the dataset, this process is known as data cleaning. The missing values can be filled and the noise can be eliminated by using some techniques like filling with most common value in missing place. The categorical variables, if any, need to be converted to dummy variables before the application of machine leaning models.

Exploratory Data Analysis : In this phase, the correlation between different attributes has to computed to see which attributes contribute significantly to the disease. Further exploration has to be done to calculate various statistical parameters like mean, median, standard deviation for various important attributes to derive key insights.

Model Fitting : After the application of various supervised learning classification algorithms to the sanitized dataset, the accuracies were noted and compared to derive key insights about the best results. The accuracies achieved were as follows:

Algorithms used to carry out the analysis:

Logistic Regression: Logistic regression is a supervised learning classification algorithm used to assign classes/labels to data instances. The class label is binary in nature, having data coded as either 1 (Class A) or 0 (class B). The predictor algorithm in logistic regression uses a sigmoid function which is as follows:

$$f(x) = \frac{1}{1+e^{-x}} \quad \text{- (1)}$$

The input values are combined with linear weights as in linear regression to give an output value but in logistic regression instead of simply using the output value, the sigmoid function takes as input the value estimated by predictor variables and converts it into a probability value, since the output of the sigmoid function lies between 0 and 1. Then depending upon the threshold value used in the analysis, the value of the probability obtained from the sigmoid function is used to decide to which class the data instance belongs to.

K-Nearest Neighbours: KNN works by finding the distances between a new data instance and all the classified data points given beforehand in the labeled data, thereafter selecting the specified number of labeled examples (K) closest to the unlabelled instance, then voting for the most frequent label (in the case of classification problems) or averaging the labels (in the case of regression problems). The distance chosen can vary. One may employ Euclidean distance, cosine similarity measure, minkowsky, correlation and chi-square as distance measures of the unclassified point from its neighbours.

Naïve Bayes : Naïve Bayes algorithms is a supervised machine learning algorithm used in classification problems based on Conditional probability(Baye's theorem) approach with an assumption that all the input variables are independent of each other that is, the presence or absence of a predictor variable does not have any effect on the value of other variables. This model is particularly useful in cases wherein the number of data points available is sufficiently large but the number of attributes are less and the predictive task is to be achieved in a short span of time.

Mathematically denoting,

$$P(class|data) = \frac{(P(data\,|class\,) * P(class\,))}{P(data\,)} \quad \text{- (2)}$$

P(class | data) = (P(data | class)*P(class)) / P(data)

P(class | data) is the posterior probability of a class given a label.

P(class) is the probability of a class.

P(data | class) is the likelihood which gives the probability of a label given the class.

P(data) is the prior probability of predictor.

Support Vector Machines : Support Vector Machine (SVM) can be used for both classification and regression use cases in machine learning. In the SVM algorithm, we plot each data point as a point in n-dimensional space (where n is number of attributes in the dataset) with the value of each attribute being the value of a particular coordinate in the n-dimensional space. Then, we perform classification by finding the appropriate hyperplane that differentiates data points into various classes available for classification. Hyperplanes are decision boundaries that help in discriminating the data points. Data points falling on either side of the hyperplane can be differentiated to different classes. The dimension of the hyperplane depends upon the number of attributes. In general for an n-dimensional feature space, the hyperplane is an (n-1) degree function. Say, if the feature space has 2 attributes then the hyperplane is a line.

Decision Trees : Decision Trees fall under the category of Supervised Machine Learning algorithms where the data is continuously split depending upon the value of a certain parameter. The tree consists of two entities, which are, decision nodes and leaves. The leaves are the decisions or the final outcomes where the instance is finally classified. The decision nodes are where the data is split based on a predefined parameter. There are many algorithms which can be

substantial in constructing Decision Trees, but one of the most effective algorithm is the ID3 Algorithm (Iterative Dichotomiser 3).

Random Forests : It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a problem and to improve the performance of the model by countering the problem of overfitting. As the name suggests, in Random forests we have multiple decision trees operating in a cascaded fashion upon different subsets of the dataset. It is advantageous over decision trees in terms of predictive accuracy since multiple decision trees act on the data in synchrony and the final classification is a majority vote based out of the individual classifications of each decision tree.*

IV. RESULTS

Table 2. Predictive Accuracy scores

Algorithms Used	Accuracy achieved(in %)
Logistic Regression:	85.25
Naïve Bayes	85.25
K-Nearest Neighbours	67.21
Support Vector Machines	81.97
Decision Tree	81.97
Random Forest	95.08

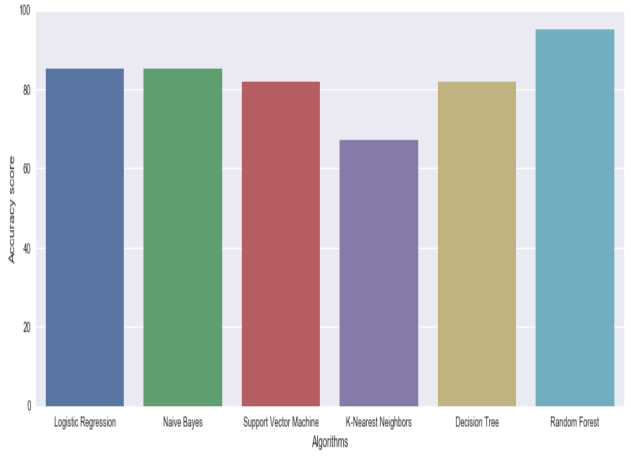

Figure 1 : Comparative visualization of algorithmic accuracies

V. Conclusion

Based on the review performed, it can be concluded that there is a huge scope for machine learning algorithms in disease prediction especially in the case of cardiovascular diseases due to the abundance of available data and extensive research done in this field. Each of the above-mentioned algorithms have performed extremely well in some configurations. Random Forest and Ensemble models have performed well because they counteract the problem of overfitting by employing an assembly of multiple decision trees. The Naïve Bayes classifier was computationally very efficient and accurate as well. SVM performed extremely well for most of the cases. This project can be substantial in serving the purpose of predicting heart ailments in individuals who interact with the system. Such kind of automation completely eliminates the scope of human error and though not a 100% precise in its working it can still be at par with the predictive ability of many certified practitioners. Integrating health and technological infrastructure in terms of disease prediction can have promising consequences in the future and lead to better care of patients due to correct prognosis.

REFERENCES

[1]. T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019

[2]. Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011

[3] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Dis. 2015;7(1):129–37.

[4] T. R. Reed, N. E. Reed, and P. Fritzson, —Heart sound analysis for symptom detection and computer-aided diagnosis,‖ Simul. Model. Pract. Theory, vol. 12, no. 2, pp. 129–146, 2004.

[5] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[6] Lamido Yahaya, Nathaniel David Oye, Etemi Joshua Garba. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques. American Journal of Artificial Intelligence. Vol. 4, No. 1, 2020, pp. 20-29. doi: 10.11648/j.ajai.20200401.12

[7] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field, in Machine Learning Paradigms, 2019, pp. 7199.