Heart disease prediction using efficient ML algorithms: A study of different techniques

Meena Rao
Dept. of ECE
Maharaja Surajmal Institute of Technology
New Delhi

Abstract

Study and prediction of heart diseases has become a critical feature of medical care in today’s world. These diseases are now one of the main causes of death in India as well as around world. Hence, there is a need for having a realistic system that can help diagnose and predict the disease at the right time. However, predicting heart diseases is not a simple task. It involves taking into account various body parameters. Hence, there is a need to apply machine learning algorithms to take into consideration various parameters and predict the heart diseases effectively. This paper studies and compares various machine learning algorithm to predict if the patient has some heart related disease or not. A total of fourteen attributes are used and four different techniques are applied for prediction. It is seen that linear regression gives the best possible output.

Keywords: ML techniques, heart disease prediction

1. Introduction

Heart is one of the major organs of our body and it working efficiently is extremely important for any healthy person. World Health Organization (WHO) estimates around 12 million deaths occur worldwide due to heart diseases [1]. Heart diseases remain the leading cause of health concern as well as mortality. Heart related diseases also increase the spending on health care and also reduce the productivity of an individual [2]. It can be safely said that prediction of heart diseases is very critical for sustainable living of any individual. As the number of patients with cardiovascular diseases are increasing, the patient load on the healthcare system is also exponentially rising. Keeping this scenario in view, it can be said that it is important to have a system in place that pinpoints or predict the heart disease accurately. The large amount of data can be easily analyzed with the machine learning algorithms. ML algorithms proves to be effective in assisting and making decisions and predictions from the large quantity of data produced by the health care industry. The main aim of this paper is to find out the likelihood of the patient to be diagnosed with any cardiovascular diseases based on their medical attributed. To do this analysis, data set has been collected from Kagel and fourteen key attributes have been considered. Four algorithms namely, Logistic Regression, Random Forest, K nearest neighbour, Logical Regression. Out of these four algorithms, Logical Regression is shown to give the best possible result.

2. **Related Work**

Research has been done in the past to predict the chances of a heart disease using ML algorithms.

An Intelligent Heart Disease Prediction System (IHDPS) developed by using data mining techniques Naive Bayes, Neural Network, and Decision Trees was proposed by Sellappan Palaniappan et al. [3]. Here, each method was shown to have its own strength to get appropriate results. To build this system hidden patterns and relationship between them were used. The technique was web-based, user friendly & expandable. In 2011, Ujma Ansari made use of Decision Tree model to predict heart disease [4]. The work utilizes dataset with 3000 instances but references of the data is not provided. Yao et al proposed the benefit and drawback of quantum clustering algorithm. They explored the improved algorithm. No pre-processing data was applied here [5]. Singh et al and Pahwa et al demonstrated multiple algorithms to diagnose heart issues with certain outcomes [6][7]. It was further seen in literature that there exist difficulties in handling large data in traditional computers. Moller and Vuik in 2017 discussed in their paper how quantum computing can lead to faster processing [8]. It was also seen in literature that heart disease dataset has to be chosen properly before applying any machine learning algorithms.

Mohan et al. had shown machine learning algorithms and neural networks are the best tools for predicting heart diseases [9]. Mc Pherson et al. also used some neural network techniques to figure out if the patient has some disease or not. It was further shown by [10] [11] that better and accurate results are obtained using machine learning algorithms.

From the literature survey, it was deduced that ML algorithms are best suited to predict heart diseases. Although work has been done in the past by various researchers but the comparisons to find out the best suited algorithm has to been done extensively. Most of the literature work focusses on two or three algorithms to deduce its suitability to predict heart diseases. In this work, authors implement four algorithms and consider fourteen attributes to find out the which algorithm is most suited to predict heart disease.

3. Algorithms used

3.1 Logistic Regression

Logistic Regression is the most often used machine learning algorithm technique. This falls under the category of supervised learning. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. Consequently, the outcome must be a known value. It can be either True or False, 0 or 1, or Yes or No. However, it provides probabilistic values that fall between 0 and 1 rather than the precise values of 0 and 1. Except for how it is applied, logistic regression and linear regression are relatively similar. The most common method for resolving classification issues is logistic regression. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

3.2 K nearest Neighbour (KNN)

One of the simplest and most effective classification approaches is KNN. This method is used when there is not much knowledge about the data distribution, even if the data is not assumed in this case. The method used in this case is to locate the K data points in the training set that are closest to the data point for which a target value is missing and then assign the average value of those data points to that data point. KNN is seen to show better results with the ant colony optimization approach. Although the K-NN approach is most frequently employed for classification problems, it can also be utilised for regression. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Algorithm:

1. Load the data
2. K is initialized to your chosen number of neighbours.
3. For the current example or chosen value, calculate the distance between the query example and the current example from the data.
4. Distance and the index of the example is added to an ordered collection
5. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
6. First K entries are to be picked from the sorted collection
7. Get the labels of the selected K entries
8. If regression, return the mean of the K labels
9. If classification, return the mode of the K labels

3.3 Random Forest Algorithm

Random Forest is a collective learning technique for classification and regression. Here, multiple decision trees are constructed for the purpose of training and finding the output of classification or regression. Following are some of the basic steps for implementing Random Forest Algorithm.

1. M bootstrap samples are randomly drawn from the training set along with replacement.
2. Grow a decision tree from the bootstrap samples.
3. Further, at each node, K features are randomly selected. This is done without replacement.
4. Next, the node is split by finding the best cut among the selected features that maximizes the information gain.
5. First two steps are repeated T times to get T tress.
6. Finally, the predictions are aggregated by different trees via the majority vote.

3.4 Linear Support Vector Classifier (SVC)-

With a high number of data, the Linear Support Vector Classifier (SVC) approach performs well. It uses a linear kernel function to perform classification. When compared to the SVC model, the Linear SVC adds more parameters including the loss function and penalty normalization, which applies "L1" or "L2." Since linear SVC is based on the kernel linear technique, the kernel method cannot be modified. Classification using linear support vectors. Both dense and sparse input are supported by this class, and the multiclass support is handled using a one-vs-the-rest strategy.

4 Methodology and Proposed Solution

The work done here involves a classification problem, so the parameters accuracy, precision, recall and F1 score are used to assess the models. In this context the terminologies: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) is used. A TN is a negative outcome predicted by the model correctly while a False Negative (FN) is a negative outcome predicted by the model incorrectly. Cross validation is not used because of insufficient dataset. In this work, dataset is split into 80% for training and 20% for test.

The system functioning begins with the collection of data as well as the significant attributes. Further the data is then preprocessed into the required format. It is then separated into training and testing parts. Training data is used for modelling the system and testing data is then used for finding the accuracy. Following modules are involved in the system: Collection of dataset, selection of attributes, data pre-processing, balancing and finally disease prediction.

4.1 Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes forthe prediction system. This is used to increase the efficiency of the system. Fourteen attributes of the patients have been considered here like age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood pressure, maximum heart rate achieved etc. and target is our target variable which has to be predicted. For attribute selection, correlation matrix has been applied.

S.No	Name	Description
1	Age	Age in Years
2	Sex	1= Male, 0= Female
3	Cp	Chest pain type (1 = typcal angina, 2= atypical angina, 3= non - anginal aymptomatic)
4	trestbps	Resting blood sugar (in mm Hg on admission to hospital)
5	Chol	Serum Cholestrol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1= true , 0 = false)
7	restecg	Resting electrocardiographic results (0= normal, 1= having ST-T wave abnormality, 2= left ventricularhypertrophy))
8	thalach	Maximum heart rate
9	exang	Exercise induced angina
10	oldpeak	ST depression induced by exercise relative to rest
11	slope	Slope of the peak exercise ST segment (1= upsloping, 2= flat, 3= downsloping)
12	Ca	Number of major vessels colored by fluoroscopy
13	thal	3= normal, 6= fixed defect, 7=reversible defect
14	Num	Class (0= healthy, 1 = have heart disease)

Fig. 3: Data Pre-processing

4.2 Prediction of Disease

Various machine learning algorithms like Logistic Regression, K nearest neighbour, Random Tree and Linear SVC are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy can then be used for heart disease prediction. Description and implementation of all the four algorithms is given in Section 3

4.3 System flowchart

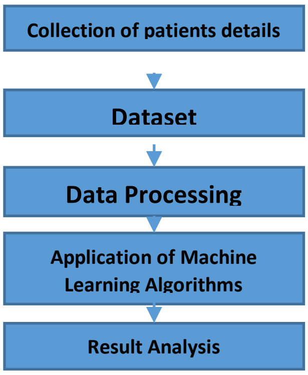
Collection of patients details
Dataset
Data Processing
Application of Machine Learning Algorithms
Result Analysis

Fig.4: System Flowchart

At first, the patient data is collected for analysis. The data set is then given attributes. Further, various machine learning algorithms are applied. Disease prediction accuracy of each algorithm can then be seen.

5. Results

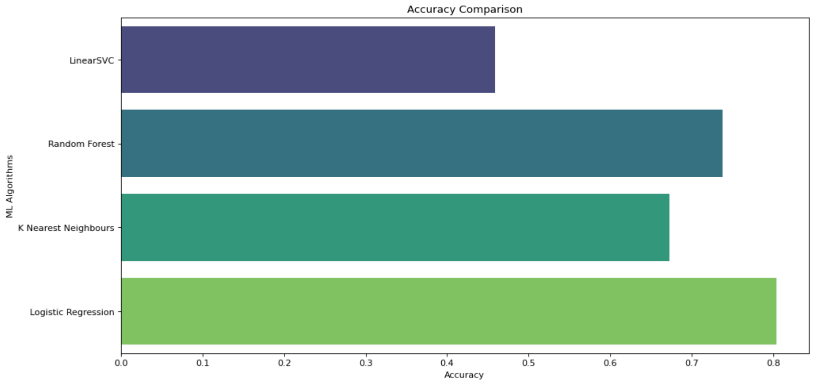

Fig. 5: Results obtained from machine learning algorithms

From the results obtained in Fig. 5, it can be observed that Logistic Regression is the most accurate in predicting heart disease. The technique shows an accuracy of 80%.

Random forest algorithm gives the next best results with an accuracy of around 75%. This is followed by accuracy of K nearest neighbours and least accuracy is shown by Linear SVC.

It can be concluded that heart diseases can be predicted by various machine learning algorithms. However, considering the various parameters, Logistic regression algorithm is seen to give best results.

References

[1] Ramalingam V.V, Dandapath A., Raja M.K, “Heart Disease Prediction using Machine Learning: A Survey”, International Journal of Engineering and Technology, Vol. 7 No. 28, pp. 684-687, 2018

[2] C. A. Devi, S. P. Rajamhoana, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika, ``Analysis of neural networks based heart disease prediction system," in Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI), Gdansk, Poland, Jul. 2018, pp. 233_239.

[3] S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques," *2008 IEEE/ACS International Conference on Computer Systems and Applications*, 2008, pp. 108-115, doi: 10.1109/AICCSA.2008.4493524.

[4] Soni J., Ansari U., Sharma D., Soni S. et al. "Predictive data mining for medical diagnosis: Anoverview of heart disease prediction, "*International Journal of Computer Applications Vol.* 17 No.8, 2011, pp. 43-48.

[5] Yao Z, Peng W, Gao-yun C, Dong-Dong C, Rui D, Yan Z(2008) Quantum clustering algorithm based on exponent measuring distance. In: 2008 IEEE international symposium on knowledge acquisition and modeling workshop, pp 436–439. IEEE

[6] M. Singh, L. M. Martins, P. Joanis and V. K. Mago, "Building a Cardiovascular Disease predictive model using Structural Equation Model & Fuzzy Cognitive Map," *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Vancouver, BC, Canada, 2016, pp. 1377-1382, doi: 10.1109/FUZZ-IEEE.2016.7737850.

[7] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Mathura, India, 2017, pp. 500-504, doi: 10.1109/UPCON.2017.8251100.

[8] Möller, M., Vuik, C. On the impact of quantum computing technology on future developments in high-performance scientific computing. *Ethics Inf Technol* Vol.**19**, pp. 253–269, 2017. https://doi.org/10.1007/s10676-017-9438-0

[9] B. S. S. Rathnayakc and G. U. Ganegoda, ‘‘Heart diseases prediction with data mining and neural network techniques,’’ in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp. 1–6.

[10] H. A. Esfahani and M. Ghazanfari, ‘‘Cardiovascular disease detection using a new ensemble classifier,’’ in Proc. IEEE 4th Int. Conf. Knowl.- Based Eng. Innov. (KBEI), pp. 1011–1014 Dec. 2017.

[11] F. Dammak, L. Baccour, and A. M. Alimi, ‘‘The impact of criterion weights techniques in TOPSIS method of multi-criteria decision making in crisp and intuitionistic fuzzy domains,’’ in Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE), vol. 9, pp. 1–8, Aug. 2015