

Heart Disease Prediction Using Machine Learning

1st Poonam Sathe

Computer Engineering

Dr.D.Y.Patil Institute of Technology

Pimpri, Pune.

2nd Renuka Sirsat

Computer Engineering

Dr.D.Y.Patil Institute of Technology

Pimpri, Pune

3rd Tejas Varude

Computer Engineering

Dr.D.Y.Patil Institute of Technology

Pimpri, Pune.

4th Yash Wandile

Computer Engineering

Dr.D.Y.Patil Institute of Technology

Pimpri, Pune

Prof. Minal Shahakar.

Dr.D.Y.Patil Institute of Technology

Pimpri, Pune.

Abstract

It happens so many times that you or someone need doctors' help but they are not available due to some reason. The Heart Disease Prediction application is a customer or end-user support to the online. In this System propose a web application that allows customer or user to get instant solution on their heart disease through an intelligent system online. The application is used with various details and the heart disease associated with those details. The applications allow to customer share their heart-related problems. It then processes user specific details to check for various illnesses that could be associated with it. Here in this system use some intelligent data mining techniques to the most accurate result find and give the solution based on patient's details. Based on the result, the automatic system shows the specific doctors for further treatment. And the system allows customers or users to view doctor's details.

Index Terms

Machine learning Algorithm, Logistic Regression, Decision Tree, Naïve Bayes, Confusion Matrix, Random Forest

1. Introduction

In today's world, heart disease is the reason for the occurrence of many deaths. Heart disease may occur due to inappropriate lifestyle, any workload, tension, pressures, and high amount of taking smoking alcohol fat which may cause hypertension. According to the World Health Organization (WHO), millions of people are dying due to Heart diseases every single year around the world. The main challenging thing in today's healthcare is the provision of the best quality of services and effective accurate diagnosis which can make little more change in the current situation. Machine Learning techniques such as Naive Bayes, Decision Trees, Logistic Regression, and Random Forest are used to predict the chances of heart disease; Machine Learning is an application of Artificial Intelligence that uses different analytics and statistical problems in order to improve the performance of particular machine learning from old data. It enables a particular machine to learn from a database and hence the performance by experience. The system used the patient details and find the patient had any disease or not. In this system provides the level of Heart Disease presence such as no HD, in this used the different models one is patient and another doctor model project will solve by two classification algorithms of supervised Machine Learning such as Decision Tree and Naive Bayes. In this systems have one admin user that manages and controls all system and data of doctors and patients reports. This project work makes an attempt to detect these heart diseases at an early stage and give the best solution to prevent heart disease, Data mining techniques use extracting valuable data and hidden information from a large amount to convert small format mostly the medical database consists of certain information which considered as discrete. Hence, making decisions using discrete data becomes a complicated and flinty task. The main goal of this system is to provide a tool for doctors to detect heart disease as early-stage. All these techniques mostly cause imprecise diagnosis and often delay in the diagnosis results due to human errors. It is more expensive and computationally complicated and takes time in assessments designed an expert medical diagnosing heart disease system and applied machine learning techniques such as NB, DT, and ANN in the system. The Naive Bayes predictive model obtained a performance accuracy of 87.16 the second model was ANN which obtained an accuracy of 87.10%, and the decision tree classifier achieved 82.40% with accurate prediction.

II. PROPOSED MODEL

In this System use the different algorithm to find the accurate result, in proposed system use the different modules one is a doctor and another is patient, inpatient module used the all patient information like name, contact number, address, Date admitted or discharge stored in the database, in doctor model used all doctor information like Doctor name, Address, in specialist in which field.

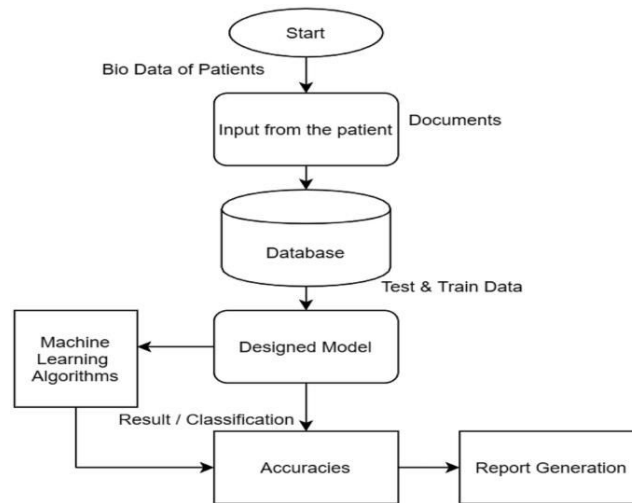


Fig1 flowchart of Heart Disease model

This proposed system stored data in the database and designed the model using machine learning algorithm in this system use UCI database .dataset using 76 attributes but in our project use in 14 structure and unstructured attribute. Classified the patient information and check the disease using all algorithms classified the disease and give the accurate result and generate the patient report the patient is a disease or not, using binary value 0 or 1.first stored all patient bio data in database divide the data in two model training and dataset and design the model using machine learning algorithm and find the accurate result of patient disease and generate a patient report

How does the Model work?

Figure 2 Steps carried out during the prediction of heart disease.

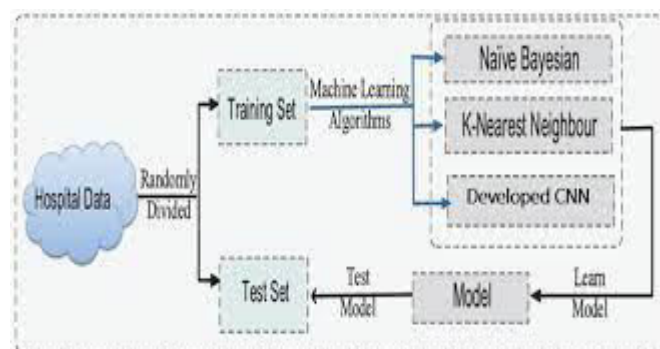


Fig2.Proposed prediction heart disease of mode

1. It starts with collect the patient information;

In these first steps, different data types used, mainly structured, or unstructured can be collected from different sources like hospitals, etc.

2. Once the information is collected, and then divides the information into two different types one is the training set and the test set

3. After that, studies of different algorithms can be used in machine learning to classify heart disease. The research was used to Naïve Bayes, K-nearest algorithm and developed CNN algorithm for classification. This process mainly used training data to find the machine efficiency and accuracy

4. Once the filled information is correct the model will be ready for testing.

5 Test all information using all algorithms and find the efficiency and accuracy of the model

6. If the predictive efficiency and accuracy are up to the desired level, then the model is deployed. And give the correct result

Data-set and Mode

Hospital Data

The medical datasets in this system used the hospital data and this data stored in our project datasets. There are 14 attributes used in the project. The structure data is considered as patient details and used in laboratory data. Such as Name, Age, Contact Details, Address, Gender this all data is used to find the relevant diagnose of heart disease inpatient. Use for future scope, unstructured data considered also used in future

Disease Risk Prediction

In this System, the main focus is completely on the patient heart disease Prediction. Here, the system requires the separate input values that can relevant to the attribute values of the patient ($a_1, a_2, a_3, \dots, a_n$) the purpose of the model design is to predict the heart disease of the patient.

Data Pre-processing

The dataset is divided into two types training data and test data to generate accurate value to evaluate the system design. Loss of the data due to errors of humans and systems etc. Thus to maintain efficiency and accuracy of the system, we need to fill in the information before data are applied to the model, loss or missing attributes filled in the system, and remaining data not needed to add. UCI database use 76 attributes but our project using 14 attribute

Structured and unstructured datasets

Structured Data in the structured dataset used all patient-related data and find the solution such as Patient Name, Age, and Address, admitted Date, Height, Blood Group, Weight, and Blood Pressure range

Unstructured Data find the patient disease infection in the body and give the solution. Stored all doctor information, records

B. Classification:-

The attributes mentioned in the project such as Confusion Matrix, Decision Tree, Logistic Regression algorithm for each of the Algorithm used in machine learning technique. The concept is evaluated to the different metrics used as exactness. Recall F-measure used to find the different accuracy results. In this system, an input dataset is used to break of the 80% training dataset and 20% remaining test dataset.

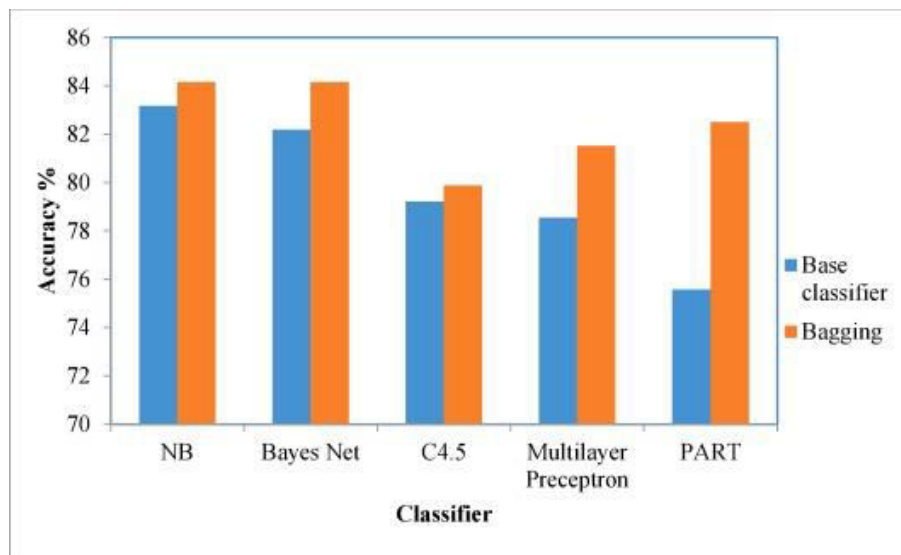


Fig 3 classification of all algorithms

1. Random Forest:-

Random forest algorithms used for regression and classification processes. There are two stages one create an algorithm and after the predication using a classifier

2. Decision Tree:-

Decision tree used for a flowchart. There are used outer branches where they represent outcomes and dataset attributes represent inner node

3. Logistic Regression:-

Logistic Regression algorithm used for binary classification problem. in this algorithm use a hyper plane or fitting a straight line. Using a logistic function to squeeze output in a line linear equation between value 1 or 0

4. Naive Bayes:

These algorithms based on the Bayes theorem. It works maverick for binary and multi-class algorithms. The main supposition of this is building a classification using family algorithms. Bayes rule computes the probability of an event.

$$P(A|B) = (P(B|A) P(A)) / P(B)$$

Table I

Classifications of accuracy and usage frequency of different supervised ML algorithms for Heart Disease prediction.

Supervised ML algorithms	Precision	Recall	Measure	Accuracy
Decision tree (DT)	0.88	0.80	0.80	81.96%
K-nearest neighbour	0.81	0.81	0.81	83.32%
Logistic regression	0.86	0.85	0.88	81.95%
Naive Bayes	0.87	0.92	0.82	80.23%
Random forest	0.79	0.86	0.89	82.13%

Table II

To get value for confusion matrix using various algorithm.

Algorithm	True (TP)	Positive (FP)	False (FN)	Positive (TN)	False (FN)	Negative (TN)
Decision tree (DT)	23		3		6	30
Random forest	20		4		6	28
Logistic Regression	21		7		4	31
Naive Bayes	24		6		3	30

Table III- Confusion matrix:-

Predicted Heart disease patient (1) Predicted healthy person (0) Actual Heart disease patient (1) TP/FN Actual healthy person (0) FP (1)

	Heart disease patient (1)	healthy patient (0)
Actual Heart disease patient (1) TP /FN	TP	FN
Actual healthy person (0) FP/TN	FP	TN

Classification accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%. \quad (1)$$

$$\text{Classification error} = \frac{FP + FN}{TP + TN + FP + FN} \times 100\%. \quad (2)$$

$$\text{True positive rate: TP} = \frac{TP}{TP + FN} \times 100\%. \quad (3)$$

Specificity: when diagnostic test is negative and the person is healthy. Calculate mathematically.

$$SP = \frac{TN}{TN + FP} \times 100\%. \quad (4)$$

The equations of precision are given as follows.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \times 100\% \quad (5)$$

$$\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN}) \quad (6)$$

MCC value is near to 0 means the classifier generates random predictions .the mathematical equation of MCC is as following.

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{(\text{TP} + \text{FP}) (\text{TP} + \text{FN}) (\text{TN} + \text{FP}) (\text{TN} + \text{FN})} \times 100\% \quad (7)$$

$$\text{F - Measure} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

- 1) TP True positive: - The test is positive and the patient is a disease.
- 2) FP False positive: - The test is not positive and the patient is not a disease.
- 3) TN True negative: - The test is negative and the patient is not a disease.
- 4) FN False negative: - The test is negative and person is the disease.

III. Experimental Result and Discussion

In this system, we discussion demand the model classification and outputs on different outlooks. First, we study the execution from different Machine Learning algorithms such as decision tree k-nearest neighbour Naive Bayes, logistic Regression, and confusion matrix on the Cleveland heart disease dataset on full function. The 6 most important features selected in table I, the decision tree show performance has 81%. The Naive Bayes algorithm shows performance is 80% accuracy

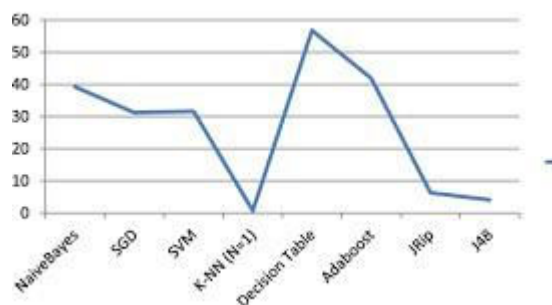


Fig 4:AUC of different classification

The random forest algorithm shows 84% performance Accuracy.AUC used for both testing data and training data SVM was used for the 85% to 86% accuracy in fig 5 is used a different algorithm to compare the accuracy of the algorithm. in all comparing algorithms, the Naïve Bayes algorithm in the following dig all algorithm comparing the CNN is higher accuracy than the other two algorithms. Then Naïve Bayes used 80% accuracy the KNN is 74% have

least accuracy model. The CNN algorithm is used for both unstructured and structured data for testing and show the correct result. It used a very good efficient algorithm to increase the accuracy of the CNN algorithm

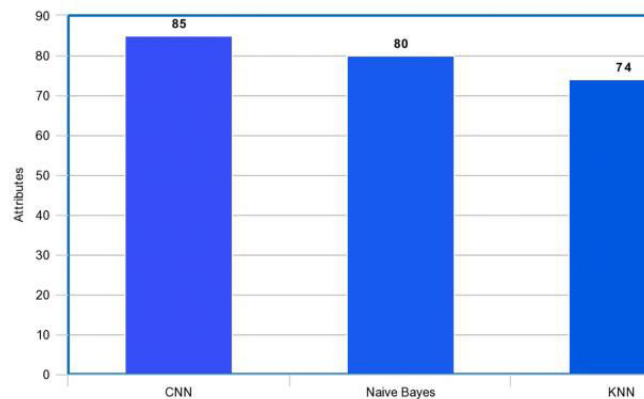


Fig 5. Compare Graphs of different algorithm

4. CONCLUSION

In the world number of deaths due to heart disease. I have to need to develop a system to divine heart disease exactly and successfully. In our paper simulation to study the complete performance of different supervised Machine learning algorithms in disease prediction systems. Since research cope and medical data vary widely between diseases prediction study. The comparison was the only possible thing on a common specification to the dataset and scope is entrenched, therefore, we have to select the multiple machine learning algorithm techniques on the same basic data and prediction for disease collation in the project we use various algorithm to get correctly inconstancy and execution. This survey will help to show the leading result

5. REFERENCE

1. Theresa Prince R, J. Thomas,' Human heart disease Prediction System using Data Mining Techniques International Conference on Circuit Power and Computing Technologies, Bangalore 2016.
2. Niagara M Lutimath, Chethan C, Basavaraj S Pol.,' Prediction Of Heart Disease using Machine Learning, International Journal Of Recent Technology and Engineering,(2S10), pp. 474-477, 2019.

3. UCI, Heart Disease Data Set.[Online]. Available (Accessed on May 1, 2020): <https://www.kaggle.com/ronitf/heart-disease-uci>.
4. Fahd Saleh Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal, Vol. 10, No. 6, 2019. Jan
5. Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin,"Design And Implementation Of Heart Disease Prediction Using Bayesian", International Conference on Trends in Electronics and Information (ICOEI 2019).
6. Davis DA, Chawla NV, Christakis NA, Barabási A-L. Time to CARE: a collaborative engine for practical disease prediction. *Data Min Knowl Disc.* 2010; 20 (3): 388–415. 6.
7. Sayad AT, Halkarnikar PP. Diagnosis of heart disease using a neural network approach. *Int J Adv Sci Eng Technol.* 2014; 2:88–92.
8. S. Nazir, S. Shahzad, and L. Septem Riza, "Birthmark-based software classification using rough sets," *Arabian Journal for Science and Engineering*, vol. 42, no. 2, pp. 859–871, 2017
9. Miranda E, Irwansyah E, Amelia AY, Kom S, Maribondang MM, Kom S, Salim M, Kom S. Detection of cardiovascular disease risk's level for adults using Naive Bayes classifier. *Health Inf Res.* 2016; 22 (3): 196–205
10. Joshi, S. a. N. M., 2015. Prediction of Heart Disease Using Classification Based Data Mining Techniques. *Computational Intelligence in Data Mining-Volume 2*, pp. 503-511.
11. A. M. De Silva and P. H. W. Leong, *Grammar-Based Feature Generation for Time-Series Prediction*, Springer, Berlin, Ger. Many, 2015.
12. Shined Swati B., P. A., 2015. Decision Support System on Prediction of Heart Disease Using Data Mining Techniques. *International Journal of Engineering Research and General Science*, 3 (2), pp. 1453-1458.
13. M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for pre-processing the data set for predicting fertility success rate," *International Journal of Control @eory and Applications*, vol. 9, pp. 256–260, 2016
14. O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang, and G. Li, "An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction," *Expert Systems with Applications*, vol. 68, pp. 163–172, 2017.
15. R. J. Urbanowicz, M. L. Olson, W. H. Meeker, W. S. Olson, and J. H. Moore, "Relief-based feature selection: introduction and review," *Journal of Biomedical Informatics*, vol. Cava, R. 85, pp. 189– 203, 2018