Available online at : http://www.advancedscientificjournal.com
http://www.krishmapublication.com
IJMASRI, Vol. 3, issue 1, pp. 831- 838, Jan. -2023
https://doi.org/10.53633/ijmasri

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY ADVANCED SCIENTIFIC RESEARCH AND INNOVATION (IJMASRI)

ISSN: 2582-9130 **IBI IMPACT FACTOR 1.5** **DOI: 10.53633/IJMASRI**

RESEARCH ARTICLE

HEART DISEASE PREDICTION USING MACHINE LEARNING

Nishika Mittal[1] Nikhil Verma[2] and Varun Goel[3]

[1,2,3] *Department of Information Technology, Maharaja Agrasen Institute of Technology, Rohini, Delhi*
nishikamittal0905@gmail.com , nikhil.kops@gmail.com, varungoel.cs@gmail.com

Abstract

Heart disease is considered one of the major causes of death throughout the world. It cannot be easily predicted by medical practitioners as it is a difficult task that demands expertise and higher knowledge for prediction. The research paper mainly intends to predict the occurrence of a disease based on data gathered from Kaggle and Cleveland foundation medical research, particularly in Heart Disease. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression, Random Forest Classifier, AdaBoost Classifier, and KNN to predict and classify the patient with heart disease. The strength of the proposed model was quite satisfying and was able to predict evidence of having heart disease in a particular individual by using AdaBoost Classifier which showed good accuracy in comparison to the previously used classifier such as KNN, Random forest, etc.

Keywords: Supervised, unsupervised, reinforced, linear regression, AdaBoost Classifier decision tree, python programming, jupiter Notebook, confusion matrix.

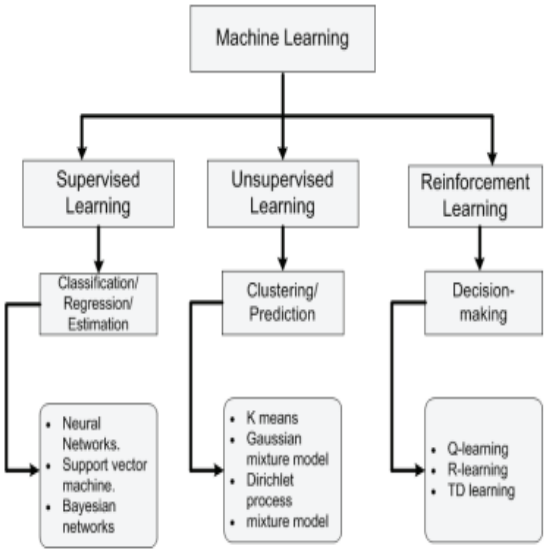
Introduction

The highest mortality of both India and abroad is mainly because of heart disease. According to World Health Organization (WHO), heart-related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. Hence, this is a vital time to check this death rate by identifying the disease correctly in the initial stage. We can use data mining technologies to discover knowledge from the datasets. Healthcare administrators can use the discovered knowledge to improve the quality of service. Anticipating patients' future behaviour on a given history is one of the important applications of data mining techniques that can be used in healthcare management.

Machine Learning is one of the efficient technologies for testing, which is based on training and testing. It is the branch of Artificial Intelligence (AI) that is one of a broad area of learning where

machines emulate human abilities. On the other hand, machine learning systems are trained to learn how to process and make use of data hence the combination of both technologies is also called Machine Intelligence.

Types of Machine Learning are:
i. Supervised Learning
ii. Unsupervised Learning
iii. Reinforcement Learning



A. *Supervised Learning*

Supervised learning can be defined as learning with the proper guide or you can say that learning in the present of teacher. we have a training dataset which act as the teacher for prediction on the given dataset that is for testing a data there are always a training dataset. Supervised learning is based on "train me" concept. Supervised learning has following processes:

- Classification
- Random Forest
- Decision tree
- Regression

To recognize patterns and measures probability of uninterruptable outcomes, is phenomenon of regression. System have ability to identify numbers, their values and grouping sense of numbers which means width and height, etc. There are following supervised machine learning algorithms:

- Linear Regression
- Logistical Regression
- Support Vector Machines (SVM)
- Neural Networks
- Random Forest
- Gradient Boosted Trees
- Decision Trees
- Naive Bayes

B. *Unsupervised Learning*

Unsupervised learning can be defined as the learning without a guidance which in Unsupervised learning there are no teacher are guiding. In Unsupervised learning when a dataset is given it automatically work on the dataset and find the pattern and relationship between them and according to the created relationships, when new data is given it classify them and store in one of them relation. Unsupervised learning is based on" self-sufficient" concept.

For example, suppose there are combination fruits mango, banana and apple and when Unsupervised learning is applied it classify them in three different clusters on the basis if their relation with each other and when a new data is given it automatically send it to one of the clusters. Supervisor learning say there are mango, banana and apple but un super vised learning said it as there are three different clusters. Unsupervised algorithms have following process:

- Dimensionality
- Clustering

There are following unsupervised machine learning algorithms: • t-SNE

- k-means clustering
- PCA

C. *Reinforcement*

Reinforced learning is the agent ability to interact with the environment and find out the outcome. It is based on "hit and trial" concept. In reinforced learning each agent is awarded with positive and negative points and on the basis of

positive points reinforced learning give the dataset output that is on the basis of positive awards it trained and on the basis of this training perform the testing on datasets.

Machine Learning recognizes who has any symptoms of heart disease such as chest pain or high blood pressure and based on these; a comparison is done in terms of the accuracy of algorithms in this pro ct we have used four algorithms which are Random forest, Logistic regression, K- neighbour, Adaboost Classifier.

In this paper, we calculate the accuracy of four different machine learning approaches and based on the calculation, we conclude which one is the best among them. A dataset is selected from the UCI repository with patient's medical history and attributes. By using this dataset, we predict whether the patient can have heart disease or not. To predict this, we use 14 medical attributes of a patient and classify them if the patient is likely to have heart disease. These medical attributes are trained under three algorithms: Logistic regression, KNN and Random Forest Classifier. The most efficient of these algorithms is AdaBoost Classifier which gives us an accuracy of 90.16**%**. And, finally, we classify patients that are at risk of getting heart disease or not.

The further paper is divided into sections, Section 1 of this paper consists of the introduction to the machine learning and heart diseases. Section II illustrated the related work of researchers. Section III is about the methodology used for this prediction system and algorithms used in this project. Section IV briefly describes the dataset and their analysis with the result of this project. And the last Section V concludes the summary of this paper with slight view about future scope of this paper.

Related work

Heart is one of the core organs of human body, it plays crucial role on blood pumping in human body which is as essential as the oxygen for human body so there is always need of protection of it, this is one of the big reasons for the researchers to work on this. So, there are number of researchers working on it. There is always need of analysis of heart related things either diagnosis or prediction or you can say that protection of heart disease. There are various fields like artificial intelligence, machine learning, data mining that contributed on this work.

(Senthilkumar Mohan *et al*., 2019) Prediction model is introduced with different combinations of features and several known classification techniques. It produced an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

Aditi Gavhane *et al.*,2018). The dataset consists of 14 main attributes used for performing the analysis. Various promising results are achieved and are validated using accuracy and confusion matrix. Using deep learning approach, 88.2% accuracy was obtained. It was also found out that the statistical analysis is also important when a dataset is analyzed and it should have a Gaussian distribution, and then the outlier's detection is also important and a technique known as Isolation Forest is used for handling this. The difficulty which came here is that the sample size of the dataset is not large. If a large dataset is present, the results can increase very much in deep learning and ML as well. The algorithm applied by us in ANN architecture increased the accuracy which we compared with the different researchers

(Senthil kumar mohan *et al.*, 2019). A quite helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quiet satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bays etc. Most efficient of these algorithms is KNN which gives us the accuracy of 88.52%. And, finally we classify patients that are at risk of getting a

heart disease or not and also this method is totally cost efficient.

(Himanshu Sharma and Rizvi 2017). Proposed "Machine Learning Techniques for Heart Disease Prediction" in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease. To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.

Nikhil Kumar, *et al.*, 2019). Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K-Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%.

(Amandeep Kaur and Jyoti Arora, 2019) Proposed a model stated the performance of prediction for two classification models, which is analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.
(Pahulpreet Singh Kohli and Shriya Arora, 2018). Proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease.

Methodology

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:

S. No.	Attribute	Description	Type
1	Age	Patient's age (29 to 77)	Numerical
2	Sex	Gender of patient(male-1 female-0)	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure(in mm Hg on admission to hospital ,values from 94 to 200)	Numerical
5	Chol	Serum cholesterol in mg/dl, values from 126 to 564)	Numerical
6	Fbs	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	restecg	Resting electrocardiographic result (0 to 1)	Nominal
8	thalach	Maximum heart rate achieved(71 to 202)	Numerical
9	exang	Exercise included agina(1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Ca	Number of major vessels (0-3)	Numerical
13	thal	3-normal	Nominal
14	Target	1 or 0	Nominal

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data.

A. Data Collection

First step for predication system is data collection and deciding about the training and testing dataset. In this project we have used 73% training dataset and 37% dataset used as testing dataset the system.

B. Attribute Selection

Attribute of dataset are property of dataset which are used for system and for heart many attributes are

like heart bit rate of person, gender of the person, age of the person and many more shown in TABLE.1 for predication system.

C. Pre-processing of data

Pre-processing needed for achieving prestigious result from the machine learning algorithms. For example, Random forest algorithm does not support null values dataset and for this we have to manage null values from original raw data. For our project we have to convert some categorized value by dummy value means in the form of "0"and "1".

D. *Data Balancing*

Data balancing is essential for accurate result because by data balancing graph we can see that both the target classes are equal. Fig.3 represents the target classes where "0" represents with heart diseases patient and "1" represents no heart diseases patients.

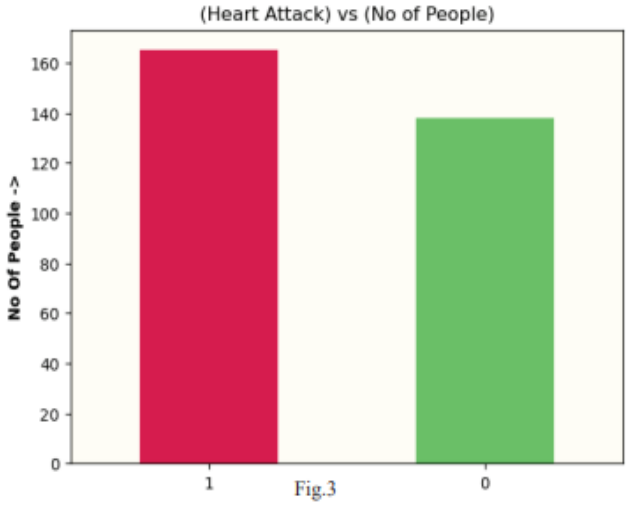


Fig.3

E. Prediction of Disease

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, K-nearest neighbor (KNN) are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

Machine Learning relies on different algorithms to solve data problems. Data scientists like to point out that there's no single one-size-fits-all type of algorithm that is best to solve a problem. The kind of algorithm employed depends on the kind of problem you wish to solve, the number of variables, the kind of model that would suit it best and so on.

A. Logistic Regression

Logistic regression is also a supervised learning classification algorithm that is used to solve both classification and regression problems. In classification problems, the target variable may be in a binary or discrete format either 0 or 1. This algorithm works on the sigmoid function, so the categorical variable results as 0 or 1, Yes or No, True or False, etc. It is a predictive analysis algorithm that works on mathematical functions.

The sigmoid functions return the value between 0 and 1. If the value less than 0.5 then it is considered as 0 and greater than 0.5 it is considered as 1. Thus, to build a model using logistic regression sigmoid function is required.

1) Binomial: The target variable can have only 2 possibilities either "0" or "1" which may represent "win" or "loss", "pass" or "fail", "true" or "false", etc.

2) Multinomial: Here, the target variable can have 3 or more possibilities that are not ordered which means it has no measure in quantity like "disease A" or "disease B" or "disease C".

3) Ordinal: In this case, the target variables deal with ordered categories. For example, a test score can be categorized as: "poor", "average", "good", and "excellent". Here, each category can be given a score like 0, 1, 2, and 3.

$$f(x) = \frac{1}{(1+e^{X})}$$

f(x) = Output between the 0 and 1 value
e = base of the natural logarithm
x = input to the function.

The value of the logistic regression must range from 0 to 1, does not go beyond this limit, so the only possible curve formed is S-shaped. The S-form curve formed is known as the sigmoid function or the logistic function. In logistic regression, the threshold value plays an important role, which defines the probability of either 0 or 1. The values above the threshold value reach 1, and a value below the threshold value reaches 0.

B. Random Forest

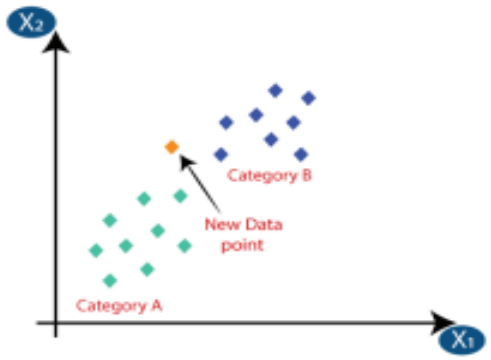
Random Forest classifier is a supervised learning technique in machine learning. It can be used to solve both Classification and Regression problems in machine learning. It is based on the process of combining multiple classifiers to solve a complex problem and to improve the performance of the model, which is known as ensemble learning. Random Forest consists of several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Rather than relying on a single decision tree, the random forest acquires the prediction from each tree, and based on the majority of votes for predictions, it predicts the final output. The higher number of trees in the forest leads to better accuracy and also prevents the problem of over fitting. The final output is taken by using the majority voting classifier for a classification problem while in the case of a regression problem the final output is the mean of all the outputs.

C. Ada-boost

AdaBoost is short for Adaptive Boosting and is a widely accepted boosting technique that combines multiple weak classifiers to build a strong classifier. It is done by building a model using a series of weak models. AdaBoost was developed for binary classification. It selects a training subset randomly and builds a model. It then iteratively trains the AdaBoost machine learning model by selecting the training set based on the accurate prediction of the last training. It assigns the higher weight to the erroneously classified observations so that in the next iteration these observations would have a higher prospect for classification. This is done to correct the errors present in the first model. It also assigns weight to the trained classifier in every iteration according to the accuracy of the classifier. The more accurate classifier will get high weight. This process iterates until the entire training data fits without any error or until it reaches the specified maximum number of models are added.

D. K-nearest Neighbor

It works on the basis of distance between the location of data and on the basis of this distinct data are classified with each other. All the other group of data are called neighbor of each other and number of neighbor are decided by the user which play very crucial role in analysis of the dataset.



In the above Fig. k=3 shows that there are three neighbor that means three different type of data are there. Each cluster represented in two dimensional space whose coordinates are represented as (Xi,Yi) where Xi is the x-axis, Y represent y- axis and i= 1,2,3,….n.

Experimental analysis

In this project, various machine learning algorithms like Random Forest, Logistic Regression,Adaboost, KNN are used to predict heart disease. Heart Disease UCI dataset, has a total of 76 attributes, out of those only 14 attributes are considered for the prediction of heart disease.

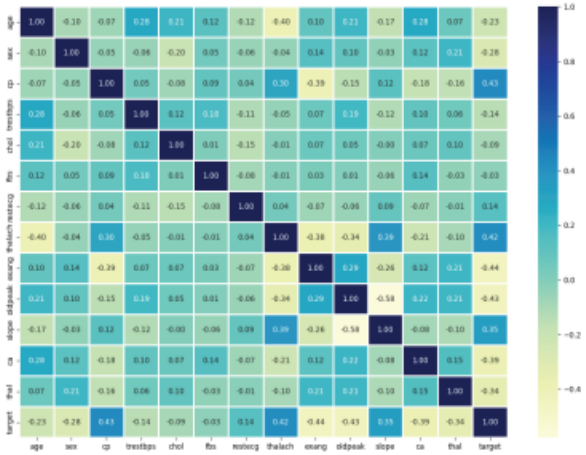
Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc. are considered for this project. The accuracy for individual algorithms has to measure and whichever algorithm is giving the best accuracy, that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered.

Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset.
It is expressed as:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.

Correlation Matrix: The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.



Precision: It is the ratio of correct positive results to the total number of positive results predicted by the system.
It is expressed as:

$$\text{Precision (P)} = \frac{TP}{(TP + FP)}$$

Recall It is the ratio of correct positive results to the total number of positive results predicted by the system.
It is expressed as:

$$\text{Recall (R)} = \frac{TP}{(TP + FN)}$$

F1 score- It is the harmonic mean of Precision and Recall. It measures the test accuracy. The range of this metric is 0 to 1.

After performing the machine learning approach for training and testing we find that accuracy of the adaboost is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm,

here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 90.16% accuracy and the comparison is shown below.

Algorithm	Accuracy
KNN	68.85 %
Random Forest	85.25 %
Logistic Regression	88.52 %
AdaBoostClassifier	**90.16 %**

Table: Accuracy comparison of algorithm

The Highest accuracy is given by AdaBoost Algorithm

Conclusion and Future scope

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the four different machine learning algorithms used to measure the performance are Random Forest, Logistic Regression, Adaptive Boosting, and K-Nearest Neighbor applied on the dataset. The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this feature have to be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account then the efficiency decreases considerably. All the four machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all four the adaptive boosting classifier gives the highest accuracy of 90%.

Reference

1. Santhana Krishnan, J and Geetha, S. (2019). Prediction of Heart Disease using Machine Learning Algorithms'' ICIICT.
2. Aditi Gavhane, Gouthami Kokkula, Isha Panday, Prof. Kailash Devadkar. (2018). "Prediction of Heart Disease using Machine Learning" Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA).
3. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva. (2019). "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" IEEE Access, 2019.
4. Himanshu Sharma and Rizvi, M.A. (2017). "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.
5. Nikhil Kumar, M., Koushik, K.V.S. Deepak, K. (2019). "Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools" International Journal of Scientific Research in Computer Science, Engineering and Information Technology, IJSRCSEIT, 2019.
6. Amandeep Kaur and Jyoti Arora, (2019). "Heart Diseases Prediction using Data Mining Techniques: A survey" International Journal of Advanced Research in Computer Science, IJARCS 2015-2019.
7. Pahulpreet Singh Kohli and Shriya Arora. (2018). "Application of Machine Learning in Diseases Prediction", 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

*****