The Heart Disease Prediction Using Bio Inspired Algorithms

[1]Pidimarla Prathyusha, [2]K.Durga Devi
[1,2]Dept. of CSE, KIET, Kakinada, AP, India

Abstract
As we all know that heart is the major organ when compared to the brain which plays an important role in the human body. It pumps blood, supplies blood to all parts of the body, and purifies the blood. Nowadays, heart diseases have become common irrespective of age. A large number of death cases all over the world are related to heart diseases. Prediction of occurrence of heart diseases is an urge nowadays the cases are increasing and heart disease does not occur all of a sudden. To deal with this problem, we need to bring about awareness about the diseases to the world. Heart diseases can be predicted with the help of Machine Learning Algorithms. A huge number of patients details will be collected and interpreted to predict the occurrence of disease. In this paper, we calculate the accuracy of machine learning algorithms for predicting heart disease. The algorithms which are used is Logistic Regression (LR), K Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and Naïve Bayes (NB) by using a sample dataset from Kaggle website. For implementation of Python programming, Jupyter notebook is used.

Keywords
Python Programming, Jupiter Notebook, Accuracy, Machine Learning Algorithm, Supervised, Unsupervised.

I. Introduction

The heart plays an important role in the human body by pumping blood, supplying blood to all the parts of the body, and purifying blood. When the heart does not get the required amount of blood, it results in heart failure and death. India is also having a very high rate of death due to heart diseases. The accurate and timely diagnosis of heart disease is necessary to improve the security of the heart and life.

A. Heart diseases

Rheumatic heart disease, Valvular heart disease, Hypertensive heart disease, and Cerebrovascular heart disease.

B. Symptoms

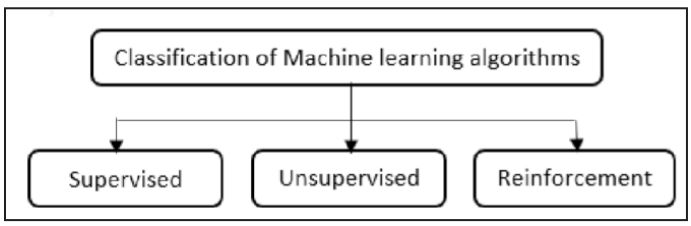
Chest pain or discomfort, shortness of breath, fainting, swelling of legs, abdomen, or areas around the eyes and easily tiring during exercise and activity.
Machine learning algorithms and techniques help us to predict and diagnose different heart diseases and help doctors to avoid sudden death in such cases. A huge number of patients details will be collected and interpreted to predict the occurrence of disease. In this project, six algorithms have been used that is Logistic Regression, K Nearest Neighbor, Decision Tree, Random Forest, Support Vector Machine, and Naïve Bayes.

C. Objective –

The objective of this project is to compare the accuracy of six different machine learning algorithms and conclude with the best algorithm among these for heart disease prediction.
Problem statement –
It is very difficult to diagnose the heart diseases in advance or at early stages even though symptoms of heart diseases are noticed and one of the common things is all of the symptoms will not occur all of a sudden. So, it's very important to diagnose the heart diseases in time which avoids death or reduce mortality rate. As we all know that diagnosing heart disease is very expensive so majority of them will not visit the doctor for consultation due to various reasons. In this paper, the problem is: based on the given information about each individual we have to calculate that whether individual will suffer from heart disease.
Justification –
In order to achieve our objective, we use machine learning algorithms at different levels of evaluation. Although machine learning algorithms are commonly used, heart disease prediction is a vital task involving the highest possible accuracy. Hence, these algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researches and medical practitioners to establish a better understanding and help them identify a solution to identify the best methods for predicting heart disease.
The main contribution of this project includes: Extraction of classified accuracy useful for heart disease prediction, comparison of different machine learning algorithms and identifying the best performance-based algorithm for heart disease prediction.
This paper consists of VII Sections. Section I deals with Introduction, Section II describes Machine Learning, Section III deals with Literature Review, Section IV is about Methodology, Section V is about Machine Learning algorithms, Section VI describes Result analysis, and Section VII is about the conclusion and future scope of this paper.

II. Machinelearning

Machine Learning is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine Learning is generally where humans train the machine and the work can be done much faster.
Machine Learning algorithms are classified into three types:



A. Supervised Learning

Supervised learning has the presence of a supervisor as a teacher. Basically, supervised learning is when we teach or train the machine using data that is well labelled. Supervised learning deals with or learns with "labelled" data. This implies that some data is already tagged with the correct answer.

Classified into two types:

* Classification – used when the output is categorical like 'Yes' or 'No'.

Algorithms used:

- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes
- K Nearest Neighbour

* Regression – used when a value needs to be predicted like the 'stock prices.

Algorithm used: Linear Regression

B. Unsupervised Learning

Unsupervised learning is the training of a machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance. Unlike supervised learning, no teacher is provided that means no training will be given to the machine.

Classified into two types:

- Clustering – used when the data needs to be organized to find patterns in the case of 'product recommendation'.

Algorithm used: K means clustering

- Dimensionality Reduction

C. Reinforcement Learning

Reinforcement learning is a feedback-based learning technique in which a machine learns to behave in an environment by performing the actions and seeing the results of actions. For each good action, the machine gets positive feedback, and for each bad action, the agent gets negative feedback or penalty.

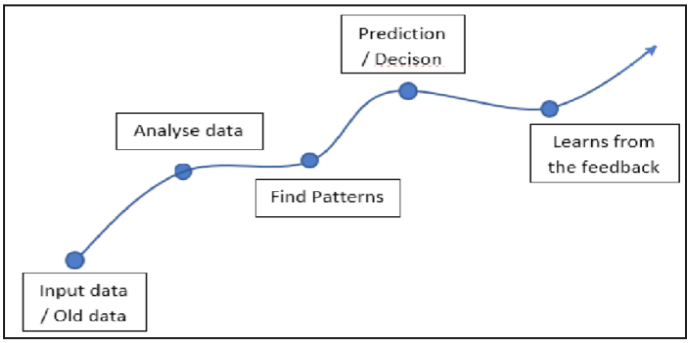
In Reinforcement learning, the machine learns automatically using feedbacks without any labelled data, unlike supervised learning.

Algorithms are not the types of machine learning. In simple words, they are the methods of solving a problem.

Algorithms to be used depends on:

- The problem statements.
- The size, quality and nature of the data.
- Complexity of the algorithm.

Machine Learning Process:



III. Literature Review

The heart is one of the core organs of the human body, it plays a crucial role in blood pumping in the human body which is an essential as the oxygen for the human body so there is always a need for protection of it, this is one of the big reasons for the researchers to work on this. So, there are a number of researchers working on it. There is always a need for 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 to this work.

Kumar et al. have worked on various machine learning and data mining algorithms and analysis of these algorithms are trained by UCI machine learning dataset which have 303 samples with 14 input feature and found svm is best among them, here other different algorithms are naivybayes, knn and decision tree.

Gavhane et al. have worked on the multi layer perceptron model for the prediction of heart disease in human being and the accuracy of the algorithm using CAD technology. If the number of person using the prediction then the awareness about the diseases is also going to increases and it make reduction in the death rate of heart patient.

Some researchers have work on one or two algorithm for predication diseases. Krishnan et al. proved that decision tree is more accurate as compare to the naïve bayes classification algorithm in their project.

Machine learning algorithm are used for various type of diseases prediction and many of the researchers have work on this like kohali et al. worked on heart disease prediction using logistic regression, diabetes prediction using support vector machine, breast cancer prediction using Adaboot classifier and concluded that the logistic regression give the accuracy of 87.1%, support vector machine give the accuracy of 85.71%, Adaboot classifier gives the accuracy up to 98.57% which good for prediction point of view.

A lot of work has been carried out to predict heart disease using the UCI Machine Learning dataset. Different levels of accuracy have been attained using various data mining techniques which are explained as follows.

AvinashGolandeet. al.; study various different ML algorithms that can be used for the classification of heart disease. The research was carried out to study Decision Tree, KNN, and K-Means algorithms that can be used for classification, and their accuracy was compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by a combination of different techniques and parameter tuning.

T.Nagamani, et al. have proposed a system which deployed data mining techniques along with the MapReduce algorithm. The accuracy obtained according to this paper for the 45 instances of testing set, was greater than the accuracy obtained using conventional fuzzy artificial neural network. Here, the accuracy of algorithm used was improved due to use of dynamic schema and linear scaling.

Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic

Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.

Anjan Nikhil Repaka, eatl., proposed a system in that uses Naive Bayesian techniques for classification of dataset and Advanced Encryption Standard algorithm for secure data transfer for prediction of disease.

Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K- Nearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different number of attributes.

Nagaraj M Lutimath, et al., has performed the heart disease prediction using Naive Bayes classification and SVM (Support Vector Machine). The performance measures used in the analysis are Mean Absolute Error, Sum of Squared Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes.

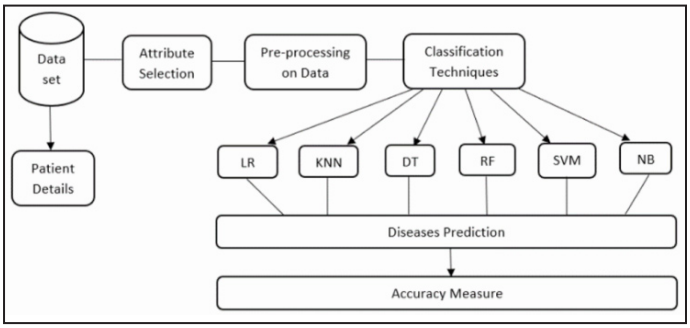
A survey paper on heart diseases prediction have proven that the old machine learning algorithms does not perform good accuracy for the prediction while hybridization perform good and give better accuracy for the predication

The main idea behind the proposed system after reviewing the above papers was to create a heart disease prediction system based on the inputs as shown. We analysed the classification algorithms namely Logistic Regression, K Nearest Neighbour, Decision Tree, Random Forest, Support Vector Machine and Naïve Bayes based on their Accuracy, Precision, Recall and fmeasure scores and identified the best classification algorithm which can be used in the heart disease prediction.

IV. Methodology

It is very difficult to diagnose the heart diseases in advance or at early stages even though symptoms of heart diseases are noticed and one of the common things is all of the symptoms will not occur all of a sudden. So, it's very important to diagnose the heart diseases in time which avoids death or reduce mortality rate. As we all know that diagnosing heart disease is very expensive so majority of them will not visit the doctor for consultation due to various reasons. In this paper, the problem is: based on the given information about each individual we have to calculate that whether individual will suffer from heart disease.

A. System Architecture:



1. Data Collection

The first step for predicting the accuracy is data collection and deciding the training and testing dataset. The dataset is taken from the Kaggle website. In this project, we have used 70% training dataset and 30% testing dataset.

2. Attribute Selection

Attribute of dataset are property of dataset which are used for system and for heart many attributes are used such as age of the person, gender of the person, heart bit rate, chest pain type and many more and also predicted output is specified in terms of 0 and 1.

Description of the attributes from the dataset is shown below:

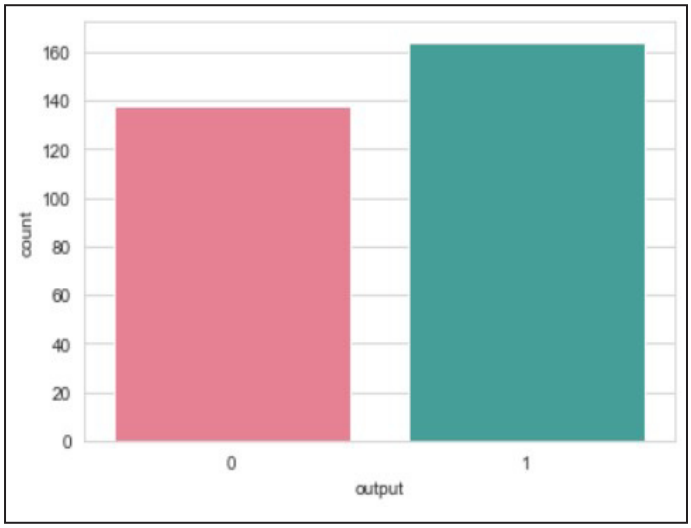
SL No.	Attribute	**Description**	**Type**
1	age	Age in years	Numeric
2	sex	Gender of the patient 1 = male; 0 = female	Nominal
3	cp	Chest pain type 0 = Typical angina: chest pain related decrease blood supply to the heart 1 = Atypical angina: chest pain not related to heart 2 = Non-anginal pain: typically non-heart related 3 = Asymptomatic: chest pain not showing signs of disease	Nominal
4	trtbps	Resting blood pressure	Numeric
5	chol	Serum cholesterol in mg/dl	Numeric
6	fbs	Fasting blood sugar > 120 mg/dl 1 = true; 0 = false	Nominal
7	restecg	Resting electrocardiographic results 0 = Nothing to note 1 = ST-T Wave abnormality 2 = Possible or definite left ventricular hypertrophy	Nominal
8	thalachh	Maximum heart rate achieved	Numeric
9	exng	Exercise-induced angina 1 = yes; 0 = no	Nominal
10	oldpeak	ST depression induced by exercise relative to rest	Numeric
11	slp	The slope of the peak exercise ST segment 0 = Upsloping: better heart rate with exercise (uncommon) = Flatsloping: minimal change (typical healthy heart) = Downsloping: signs of unhealthy heart	Nominal
12	caa	Number of major vessels (0-3) colored by flouroscopy	Numeric
13	thall	Thalium stress result 1,3 = normal 6 = fixed defect: used to be defect but ok now 7 = reversable defect: no proper blood movement when exercising	Nominal
14	output	Have disease or not 1 = yes; 0 = no (predicted attribute)	Nominal

3. Data Preprocessing

To work with categorical variables, we should break each categorical column into dummy columns with 1s and 0s. This step is one the most important step that is to be performed to get accurate result.

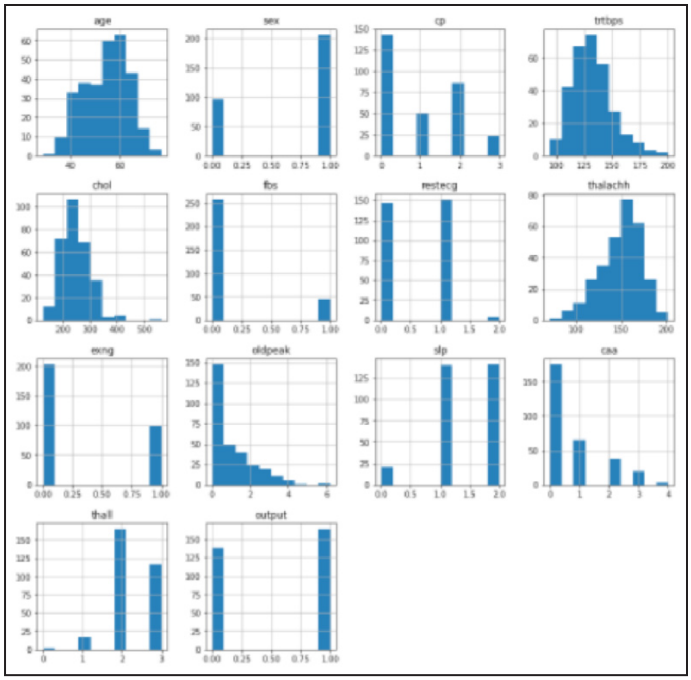
4. Data Balancing

Through data balancing we can ensure that both the output classes are balanced to move to further steps. "0" represents that the person is predicted with heart disease and "1" represents that the person is predicted without heart disease.
Below, you can see the graph:



5. Histogram of attributes

Histogram helps in understanding of each attribute clearly. The best part about this type of plot is that it just takes a single command to draw the plots and it provides so much information in return.



V. Algorithms

A. Logistic Regression

Logistic Regression is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1.

B. Decision Tree

Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to records attribute.

C. K Nearest NeighbourThe abbreviation KNN stands for "K-Nearest

Neighbour". It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

D. Random Forest

Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

E. Sector Vector Machine

A support vector machine is a supervised learning algorithm that sorts data into two categories. It is trained with a series of data already classified into two categories, building the model as it is initially trained. The task of an SVM algorithm is to determine which category a new data point belongs in. This makes SVM a kind of non-binary linear classifier.

F. Naïve Bayes

Naive Bayes algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds.

VI. Result Analysis

A. Jupyter Notebook

Jupiter notebook is used as the simulation tool and it is comfortable for python programming projects. Jupiter notebook contains rich text elements and code also, which are figures, equations, links, and many more. Because of the mix of rich text elements and code, these documents are the perfect location to bring together an analysis description, and its results, as well as, they can execute data analysis in realtime. Jupiter notebook is an open-source, web-based interactive graphics, maps, plots, visualization, and narrative text.

Algorithm	Accuracy
Logistic Regression	85%
Decision Tree	78%
K-Nearest Neighbour	62%
Random Forest	91%
Support Vector Machine	86%
Naïve Bayes	90%

B. Accuracy calculation

The accuracy of the algorithms are depends on four values namely true positive(TP), false positive(FP), true negative(TN) and false negative(FN).

Accuracy=(FN+TP)/(TP+FP+TN+FN)

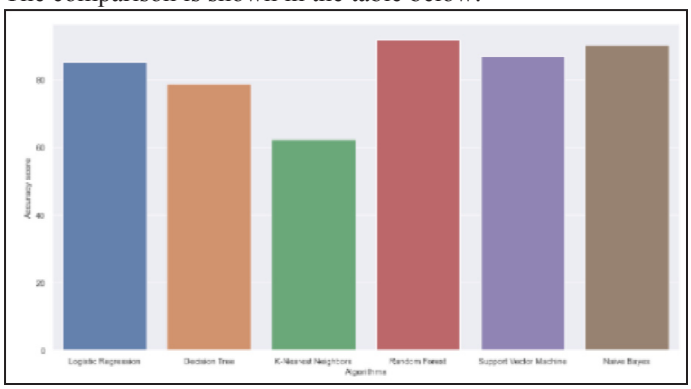
The numerical value of TP, FP, TN, FN defines as: ·TP=Number of person with heart diseases

- TN=Number of person with heart disease and no heart diseases
- FP=Number of person with no heart diseases
- FN=Number of person with no heart diseases and with heart diseases

C. Result

After performing the machine learning approach for testing and training we find that the accuracy of Random Forest is much more efficient as compared to that of the other 5 algorithms used in this project. Random forest is best among the rest with 91% accuracy.

The comparison is shown in the table below:



VII. Conclusion and Future Scope

This project provides deep insight into machine learning techniques for the classification of heart diseases. The role of a classifier is crucial in the healthcare industry so that the results can be used for predicting the treatment which can be provided to patients. With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for the detection of heart diseases. The accuracy of the algorithms in ML is dependent upon the dataset used for training and testing purposes. when we perform the analysis of algorithms on the basis of the dataset.

In the future, more machine learning approaches will be used for the best analysis of heart diseases. We can also combine two or more algorithms to form a hybrid model to get more accuracy in the future.

References

[1] AvinashGolande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] 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, Volume8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).

[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.

[6] Nagaraj M Lutimath,ChethanC,Basavaraj S Pol.,'PredictionOf Heart Disease using Machine Learning', International journal Of Recent Technology and Engineering,8,(2S10), pp 474-477, 2019