Heart disease prediction using machine learning

P. Rukmini Devi[1], Yellugari Vamshidhar Reddy[2], Yannam Satya Teja[3], Yanamala Charantej Reddy[4], Thalagaturi Venkata Sai Meenan[5]

[1]B.TECH (CSE)
[2,3,4,5] Computer Science And Engineering in Artificial Intelligence And Machine Learning Malla Reddy University, Hyderabad

ABSTRACT

Heart is the most essential part of our body. There are a lot of cases in the world related to heart diseases. People are leading to death due to heart disease. Various symptoms like chest pain, fasting of heartbeat, cholesterol etcetera. This paper gives the idea of predicting heart disease using machine learning algorithms. In order to predict much more precisely in this case, we used one of the most popular Machine Learning techniques, such as Logistic Regression. The algorithms are used on the basis of features and for predicting the heart disease. This paper uses different machine learning algorithms for comparing the accuracy among them. In order to forecast cardiac disease, the algorithms are used based on features. This study compares the accuracy of different machine learning methods.

INTRODUCTION

The human body's heart is a vital organ. It supplies blood to every organ of our body. If it doesn't work properly, the brain and several other organs will stop operating, and the individual will die in a matter of minutes. Because of numerous contributing risk factors, including diabetes, high blood pressure, high cholesterol, irregular pulse rate, and many other factors, it is challenging to diagnose heart disease.A method of manipulating and extracting implicit, previously unknown/known, and potentially relevant information in data is called machine learning . The area of machine learning is extremely broad and diversified, and its application and breadth are growing daily. Machine learning uses a variety of classifiers from supervised, unsupervised, and ensemble learning to predict outcomes and measure the accuracy of a dataset. Several techniques, including the K-Nearest Neighbour Algorithm (KNN), Decision Trees (DT), Genetic Algorithm (GA), and Naive Bayes, are used to assess the severity of the condition.Data on various health-related issues is collected by medical organisations all over the world. These data can be used to gain useful insights by utilising various machine learningtechniques.

However, the data collected is massive, and this data is frequently noisy. These datasets, which are too large for human minds to comprehend, can be explored easily using various machine learning techniques. As a result, in recent years, these algorithms have proven to be extremely useful in accurately predicting the presence or absence of heart-related diseases. The medical database is mostly made up of discrete data. As a result, making decisions with discrete data becomes a complex and difficult task. Machine Learning (ML), a subfield of Machine learning efficiently handles large scale well-formatted datasets. Machine learning can be used in the medical field to diagnose, detect, and predict various diseases. The primary goal of this paper is to provide doctors with a tool for detecting heart disease at an early stage [5]. As a result, patients will receive more effective treatment while avoiding serious consequences. ML is very important in detecting hidden discrete patterns and thus analysing the given data. Following data analysis, ML techniques aid in the prediction and early detection of heart disease.

Algorithms and Techniques

1. K – Nearest Neighbour

The k-Nearest-Neighbours (kNN) method is a simple but effective classification method. The major disadvantages of kNN are its low efficiency - being a lazy learning method precludes it from being used in many applications such as dynamic web mining for a large repository - and its reliance on the selection of a "good value" for k.
KNN gives an accuracy of 64.47368421052632%

2. Naive Bayes

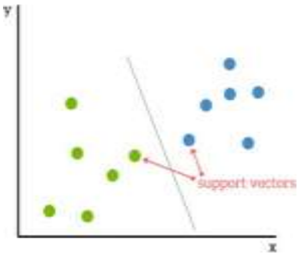
Naive Bayes is a simple but effective classification technique based on the Bayes Theorem. It assumes predictor independence, which means that the attributes or features should not be correlated or related to one another in any way. Even if there is dependency, all of these features or attributes contribute to the probability independently, resulting in why it is called Naïve Bayes.

$$\overset{\text{Posterior}}{\overset{\Downarrow}{P(A|B)}} = \frac{\overset{\text{Likelihood}}{\overset{\Downarrow}{P(B|A)}} * \overset{\text{Prior}}{\overset{\Downarrow}{P(A)}}}{\underset{\text{Evidence}}{\underset{\Uparrow}{P(B)}}}$$

Naive Bayes has achieved an accuracy of 86.8421052631579%

3. Support Vector Machine

Support Vector Machine is a well-known supervised machine learning technique that can be used as both a classifier and a predictor. It finds a hyper-plane in the feature space that distinguishes between the classes for classification. An SVM model represents the training data points as points in the feature space, mapped so that points belonging to different classes are separated by as wide a margin as possible. The test data points are then mapped into that same space and classified based on where they fall on the margin.

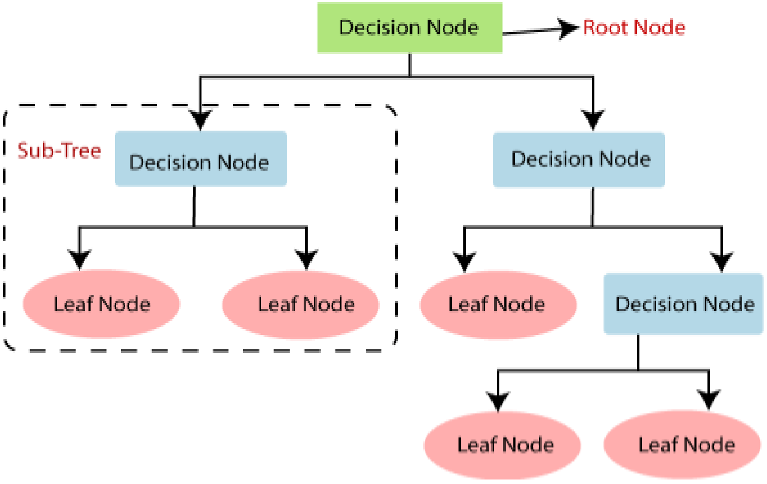


Support Vector Machine has achieved an accuracy of 89.47368421052632%

4. Decision Tree

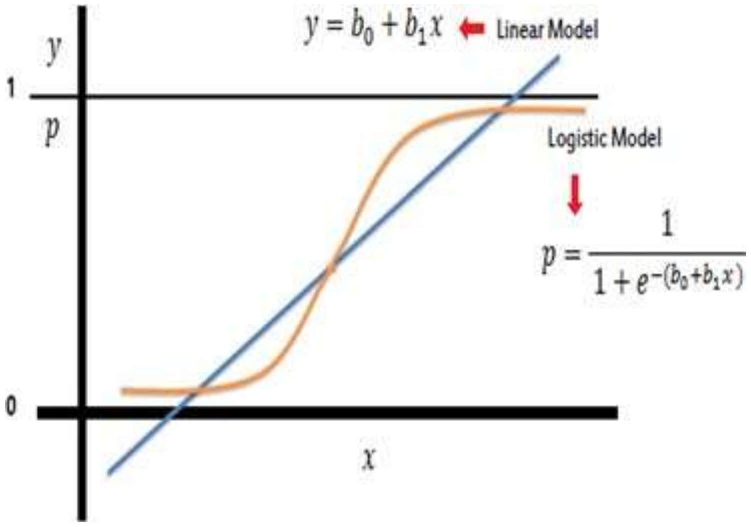
A supervised learning algorithm is a decision tree. This method is commonly used in classification problems. It works well with both continuous and categorical attributes. Based on the most significant predictors, this algorithm divides the population into two or more similar sets. The Decision Tree algorithm first computes the entropy of each attribute. The dataset is then split using the variables or predictors with the highest information gain or lowest entropy. These two steps are repeated with the remaining attributes.

The decision tree accuracy obtained is 86.8421052631579%



5. Logistic regression

Logistic Regression is a statistical and machine-learning technique classifying records of a dataset based on the values of the input fields. It predicts a dependent variable based on one or more set of independent variables to predict outcomes. It can be used both for binary classification and multi-class classification.

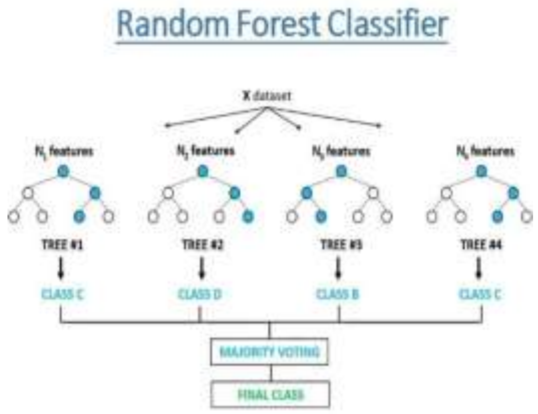


The accuracy obtained using logistic regressionis 92.10526315789474%. We achieved greater accuracy with logistic regression than with any of the other techniques.

6. Random Forest

Random Forest is another well-known supervised machine learning algorithm. This technique can be used for both regression and classification tasks, but it performs better in the latter. The Random Foresttechnique, as the name implies, takes into account multiple decision trees before producing an output.

As a result, it is essentially an ensemble of decision trees. This technique is based on the assumption that more trees will lead to the correct decision. In classification, it uses a voting system to determine the class, whereas in regression, it takes the mean of all the decision tree outputs.



Accuracy Obtained:

Algorithms	Accuracy
K– Nearest Neighbour	**64.47368421052632**
Naive Bayes	**86.8421052631579**
Support Vector Machine	**89.47368421052632**
Decision Tree	**86.8421052631579**
logistic regression	**92.10526315789474**
Random forest	**90.78947368421053**

We achieved greater accuracy of 92.10% using logistic regression than with any other algorithm. As it is extremely fast at classifying unknown records. It has good accuracy for many simple data sets and performs well when the dataset is linearly separable. Logistic regression is used to solve classification problems. Because this dataset is based on classification, logistic regression produces more accurate result of weather the person is having the heart disease or not.

Dataset:

In order to better accurately forecast the occurrence of heart disease, we used 14 variables.

These 14 attributes include:

1. Age
2. sex
3.cp
4. trestbps
5. chol
6.fbs
7. restecg
8. thalach
9. exang
10. oldpeak
11. slope
12. ca
13. thal
14. label (the predicted attribute)

Age: age in years

Sex: sex (1 = male; 0 = female)

CP: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic

Trestbps: resting blood pressure (in mm Hg on admission to the hospital)

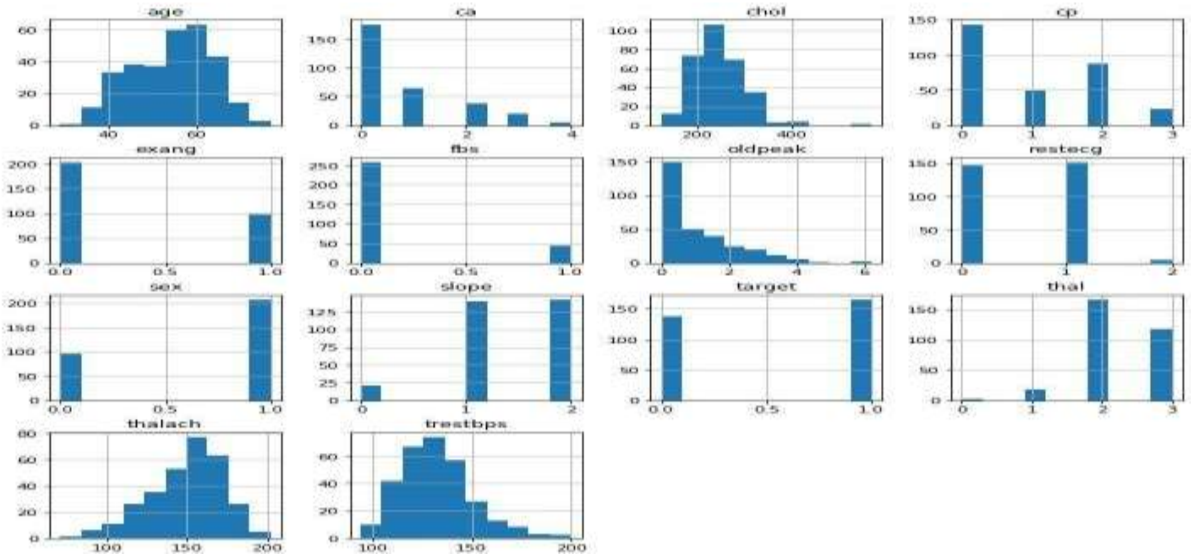
Chol: serum cholestoral in mg/dl

Fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

Restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

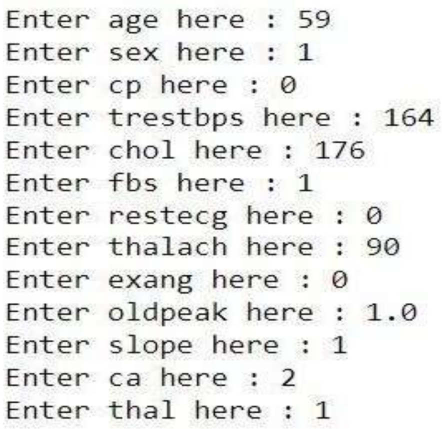
Thalach: maximum heart rate achieved
exang: exercise induced angina (1 = yes; 0 = no)
oldpeak: ST depression induced by exercise relative to rest
slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: down sloping

Ca: number of major vessels (0-3) colored by flourosopy
Thal:3 = normal; 6 = fixed defect; 7 = reversable defect



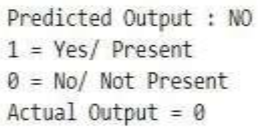
RESULT

We took a random person's data and used a machine learning algorithm to predict whether or not the person has heart disease.



```
Enter age here : 59
Enter sex here : 1
Enter cp here : 0
Enter trestbps here : 164
Enter chol here : 176
Enter fbs here : 1
Enter restecg here : 0
Enter thalach here : 90
Enter exang here : 0
Enter oldpeak here : 1.0
Enter slope here : 1
Enter ca here : 2
Enter thal here : 1
```

The machine learning algorithm will now predict whether or not the person has heart disease based on the data entered. It will predict 1 if the person has heart disease and 0 if the person does not have heart disease.



```
Predicted Output : NO
1 = Yes/ Present
0 = No/ Not Present
Actual Output = 0
```

The ML Algorithm predicted that the person does not have heart disease by using the person's data, and the predicted output matches the actual output. As a result, we can demonstrate that the algorithm for heart disease prediction is completely functional.

CONCLUSION

Heart disease is one of society's major concerns, and the numberof people affected by it is growing by the day, making it critical to find a solution. Manually calculating the chances of developing heart disease based on risk factors is difficult. However, with the help of data analytics and machine learning models, we can identify these diseases and improve our chances of curing them. Based on the preceding discussion, it is possible to conclude that machine learning algorithms have a large potential for predicting cardiovascular diseases or heart-related diseases. Among the classifiers, Logistic Regression had an accuracy rate of 92.10%, Random Forest Classifier had an accuracy rate of 90.78%, and SVM had an accuracy rate of 89.47%. To understand the classification pattern, further evaluation of the model entails generating the confusion matrix. Though all of the algorithms performed with greater than 90% accuracy, logistic regression was the most accurate.

Machine learning algorithms and techniques have been very accurate in predicting heart diseases, but there is still a lot of research to be done on how to handle high dimensional data and overfitting. A great deal of research can also be done on the best ensemble of algorithms to use for a specific type of data.

REFERENCES

[1]. https://archive.ics.uci.edu/ml/datasets/Heart+Diseas
[2]. Predicting Heart Failure: Invasive, Non-Invasive, Machine Learning, and Artificial Intelligence Based Methods.Kishor Kumar Sadasivuni (Editor), Hassen M. Ouakad (Editor), Somaya Al-Maadeed (Editor), Huseyin C. Yalcin (Editor), Issam Bait Bahadur (Editor)
[3]. Python Machine Learning - Third Edition By Sebastian Raschka, Vahid Mirjalili
[4]. Handbook of Research on Disease Prediction Through Data Analytics and Machine Learning.