International Journal of Engineering Applied Sciences and Technology, 2020
Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 489-493
Published Online June 2020 in IJEAST (http://www.ijeast.com)

RESEARCH ON MACHINE LEARNING ALGORITHM ON HEART DISEASE PREDICTION

Yash Raj Tripathi, Shruti Gupta, Shubham Mishra
Department of CSE
IMS Engineering College, Ghaziabad, Uttar Pradesh

Abstract: In today's world Heart Disease have become a very common problem even at a young age. Usually people are unware of the problem at an early stage and due to this carelessness, it proves to be a life threating problem. However, if these Heart Disease are predicted at an early stage, they can be cure.
With this inspiration, in this paper we aim for finding the best algorithm for predicting different heart conditions using different machine learning algorithms available to us.

Here, we have predicted a patient's heart condition given that we know some of the key attributes (Age, Sex, Chest Pain type etc.) of the particular person. By using different algorithms, we calculate the accuracy of each algorithm and accordingly use the best suited algorithm for the most accurate results.

Keywords: K-Nearest Neighbour, Random Forest, Gaussian, Decision tree, ANN (Artificial Neural Network using multilayer perceptron).

I. INTRODUCTION

Machine Learning prediction is basically extracting new output or information by training on or examining a huge amount of data that is collected and fed to the algorithm. All the prediction that the ML model does is based on dependent completely on the data fed to it for training.

By using machine learning algorithms with IoT (Internet of Things) devices we can easily predict diseases like Cardiovascular diseases by knowing some specific heart conditions and these methods are much more cost effective than the existing ones. These heart conditions can be predicted using several different factors that are age, sex, cholesterol level, blood sugar level, blood pressure etc. which are explained later in the paper.

- Background-Motivation:

Heart is a very crucial as well as delicate part of our body. So, it's our responsibility to look after it and ensure its smooth working. The rate of Heart Malfunctioning has increased drastically in our country and the average age at which these malfunctions occur has also decreased which is very concerning situation for all of us. This is mainly due to the busy and stressful lifestyle that we are leading.

Physical Activities have reduced and desk work has increased which makes us unfit. Another reason is unhealthy eating habits.

Most importantly, people hesitate to get themselves a heart check-up as it is very expensive, people in rural areas do not have an easy and reliable access to proper Heart check-ups.

- Objective:

The objective of this paper is to collect the data related to the heart and train the data with help of dataset called "Heart Disease Dataset". This dataset consists of four databases: Cleveland, Hungary, Switzerland and Long Beach V. It contains 14 attributes and a total of 1300 entries. This algorithm uses Supervised Machine Learning algorithm for data prediction. We will predict the probability of heart disease in a patient on the basis of given data with output as" YES (1)" and "NO (0)".

II. LITERATURE SURVEY

There are several papers that are related to the problem we are discussing. Each paper having its own approach and methodology.

In a paper, a pretty strong algorithm is used for prediction known as "Decision Tree". In this algorithm the author would combine different data sets to avoid irregularities in the model and find the best data fit for the prediction. In this only some features from the dataset are extracted in order to obtain best solution available.

In another paper the concept of feature scaling and dimensionality reduction is used for the prediction. This is shown by then applying any simple machine learning algorithm to generate output.

In another paper, the author used J48 Classifier Unpruned Tree, Decision tree, Naïve Bayes' classification and Neural Network. They used sensitivity rate for comparing the data and due to this were able to achieve a very high accuracy rate. It used the data mining model called "Transthoracic Echocardiography Report Dataset".

Another paper showed how dependencies of various attributes in any given dataset are responsible for accurate heart disease prediction. This was proved by taking 4 different databases with similar kind of attributes. The paper stated that these attributes were not enough to predict the heart disease. This was tested using different Machine Learning algorithms like Naïve Bayes', Decision Tree, Support Vector Machine, Ada Boost, MLP.

On a different paper, it reflected that the complete study of different papers that have done on heart condition and disease prediction use Data Mining Techniques and Fuzzy approach. The whole analysis is based on various factors like number of attributes, accuracy and success of the models. All the tools and algorithm used in a model are taken into consideration and then compared together to find the best possible model for this purpose.

III. PROPOSED WORK

What we have proposed here is using different Machine Learning Algorithm like Decision Tree, Naïve Bayes', K-Nearest Neighbour, Gaussian and Random Forest on the same data and finding the Accuracy that each of them provide on the given dataset.

On the basis of the found accuracy we plot the confusion matrix and find out the best suited algorithm for the current scenario and then implement it to get the desired result with maximum accuracy that is possible.

The proposed model here is to take real time data from devices like ECG and Bpm sensor and then predict that whether the patient has a Heart disease or not accordingly by using the given algorithm. This will easily identify whether the patient has heart disease or not.

For better learning of the model, we've used a large dataset. The model's learning directly depends on how rigorous the dataset is. With this we'll have an algorithm that is best suited for the job on the basis of its accuracy score.

- *Implementation strategy:*

A. Algorithm Used:

These are the algorithms that are used in the project for heart disease prediction. All of these are Supervised Learning Algorithm

In a Supervised Learning, the model is provided with a training dataset with help of which the model learns how to react in different situation with different values of the variable provided. This is what makes the base for predicting the output for any model.

- *Decision Tree:*

Decision tree algorithm is one of the most powerful tool for classification and prediction. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. As name suggests, it uses a decision tree to go from observation to conclusion about a targets' value. Tree Models where target models can take discrete values are known as Classification decision tree. Here, leaf of the tree represents class labels and branch represent conjunction of features that lead to those class labels.

Decision Tree where target variable can take continuous values are called Regression Tree.

In Decision Tree we have a process known as attributes selection with which we identify the attribute that is to be taken as the root node of the decision tree. There are two different ways of doing so-

I. Information Gain
II. Gini Index

- *K-Nearest Neighbour:*

K-Nearest Neighbour algorithm is one of the most simple machine learning algorithm. It is used for both classification and regression. In it the input consist of k training examples.

Its stores all the available data and classify new data on the basis of similarity. With this whenever a new data comes it can be classified into new category.

It does not learn from the training set immediately instead it stores the dataset and at the time of classification, because of this it is called lazy learner 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.

- *Random Forest:*

Random forest algorithm is used for both classification and regression. The major difference between random forest algorithm and decision tree is that the process of finding the root node and then splitting into it various part will run randomly.

There are two sections in Random Forest algorithm, one is random forest creation, and other one is to make a prediction from random forest classifier created in the first part.

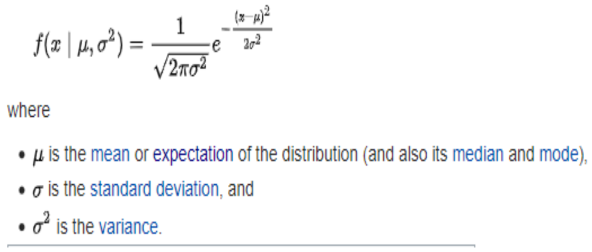
These work based on development of numerous choice trees. This will be better than choice tree on the grounds that here we are utilizing various trees to anticipate the outcomes, as there are more trees to foresee the precise outcomes. Irregular choice woods right the imperfection of choice tree calculation's propensity for over fitting to their preparation set.

It is more adaptable and easier to use machine learning algorithm that predicts and produces, even without hyper-parameter tuning. It is mostly used algorithm because of its simplicity.

- *Gaussian Classifier:*

A Gaussian classifier is a generative methodology as in it endeavours to demonstrate class back just as info class-contingent appropriation. Thusly, we can produce new examples in input space with a Gaussian classifier.

The probability density function of a unilabiate normal distribution is given by:



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where

- μ is the mean or expectation of the distribution (and also its median and mode),
- σ is the standard deviation, and
- σ^2 is the variance.

B. Data set used:

- The data set is a Heart Disease Dataset.
- This dataset consists of four databases: Cleveland, Hungary, Switzerland and Long Beach V.
- It consist of 14 attributes and 1300 data records of different persons.
- Here we are using all supervised learning algorithms for prediction.

C. The Attributes are:

These attributes are taken from the UCI machine learning repository where the attributes description is given and explained clearly.

1. Age - age in years
2. sex - sex where 1 = male; 0 = female
3. Cp - chest pain type where 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic
4. trestbps – it is a resting blood pressure (in mm Hg on admission to the hospital)
5. cho – it is a serum cholesterol in mg/dl
6. fbs – it is a fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
7. Restecg – it is a resting electrocardiographic results (0 = normal; 1 = having ST-T; 2 = hypertrophy)
8. Thalach –it is a maximum heart rate achieved
9. Exang –it is an exercise induced angina (1 = yes; 0 = no)
10. Old peak - ST sadness incited by practice comparative with rest
11. Slope - the slope of the peak exercise ST segment (1 = upsloping; 2 = flat; 3 = down sloping)
12. ca - number of major vessels
13. Thal - 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Target- the predicted attribute - diagnosis of heart

- Accuracy of each Algorithm:

Different types of algorithm like Decision Tree, K-Nearest Neighbour, Random Forest, Gaussian are used for predicting Heart condition for patients.

Python is used to follow up the work and IDE used is Spyder.

The accuracy is calculated by comparing the test data (randomly selected data from the data set) from the predicted data obtained from the algorithm.

Below is the test data from the data set

Test Data:

[*1 0 0 1 1 0 1 1 1 1 0 0 1 1 0 0 0 0 1 0 1 1 0*

1 0 0 1 0 1 1 0 1 0 1 0 0 1 1 1 1 0 0 0 1 1 0

0 1 1 1 0 0 0 0 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0

0 1 1 0 1 1 1 1 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0

1 0 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 0

***0 0 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1*]**

Decision Tree:

- Accuracy with 14 attributes

Predict Tree:

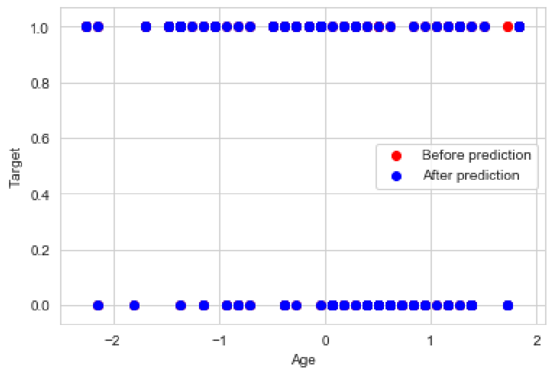
[0 0 0 1 1 0 1 1 0 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 0 0 0 1 1 1 1 0 0 1 1 1 0 0 1 1 1 0 0 0 0 0 1 1 1 0 0 0 1 1 0 1 0 1 0 0 0 1 0 1 1 1 1 1 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 0 1 0 0 1 0 1 1 1 1 0 1 0 0 0 0 1 0 1 0 1 1 1 0 1 0 1 1 1 1 1]

Accuracy: 0.8796992481203008

Final Accuracy=Accuracy*100

=87.96992481203008%

Plot Graph between Before Prediction and After Prediction:



Random Forest:

- Accuracy with 14 attributes

Predict random forest:

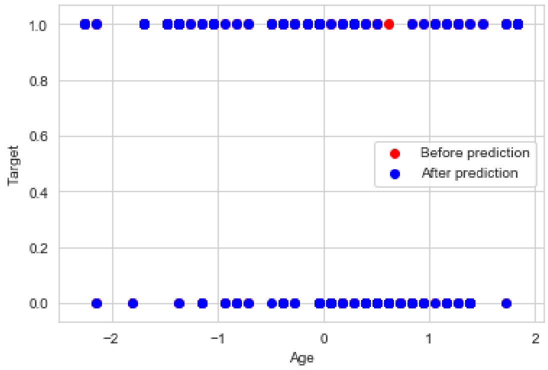
[*1 0 0 1 1 0 0 1 1 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 1 0 0 1 0 1 0 0 1 1 1 1 0 0 0 1 1 0 0 1 1 1 0 0 0 0 0 1 1 0 0 0 0 1 1 0 1 1 1 1 0 0 0 0 0 11 1 1 1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1*]

Accuracy: 0.9323308270676691

Final Accuracy=Accuracy*100

=93.23308270676691%

Plot Graph between Before Prediction and After Prediction:



Gaussian:

- Accuracy with 14 attributes

Predict Gaussian:

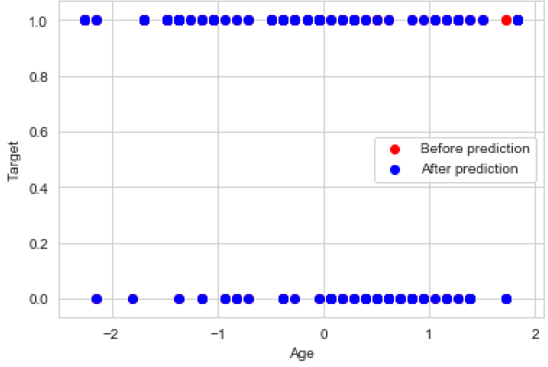
[*1 1 0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 1 0 0 0 1 1 1 0 0 0 1 1 0 0 1 1 1 0 0 1 0 0 1 1 1 0 0 0 1 1 0 1 1 1 1 0 0 1 0 1 1 0 1 1 1 0 0 0 0 0 1 0 0 0 1 0 1 1 0 1 0 1 0 1 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 0 1 1 1 0 0 0 1 1 1 1 1*]

Accuracy Gaussian: 0.8872180451127819

Final Accuracy=Accuracy*100

= 88.72180451127819%

Plot Graph between Before Prediction and After Prediction:



K—Nearest Neighbour:

- Accuracy with 14 attributes.
- Number of neighbours:5

Predict KNN:

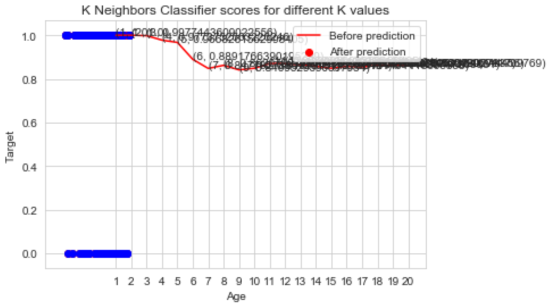
[1 0 0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 1 0 1 1 0 1 0 1 0 0 1 1 1 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0 1 1 1 0 0 1 1 0 1 1 1 1 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 0 1 0 0 0 1 1 0 1 1 1 1 1 0 0 0 1 1 1 1 1]

Accuracy: 0.9699248120300752

Final Accuracy=Accuracy*100

=96.99248120300752%

Plot Graph between Before Prediction and After Prediction:



IV. CONCLUSION

The Conclusion we get here is clearly that K-Nearest Neighbour algorithm outperforms every other algorithm by quite a margin. Its accuracy is >95% which is quite impressive. So to Predict the final Output we prefer to use K-NN algorithm as it yields the best result.

V. REFERENCES

1. Kavitha B., Kumar R.Naveen (2010): Improvising Heart Attack Prediction System using Feature Selection and Data Mining Methods ,Improvising Heart Attack Prediction System using Feature Selection and Data Mining Methods,Lecturer Department of Computer Applications, Karpagam University Coimbatore, India.
2. Krishnaiah V., Narsimha G., Chandra N. Subhash (2016): Heart Disease Prediction System Using Data Mining Techniques and Intelligent Fuzzy Approach , CSE, Research Scholar, JNTUH Dept. of CSE, Hyderabad Research Scholar, JNTUH(Pg- Vol-136).
3. Tanja Abhishek, Jain S.A.(2013): Heart Disease Prediction System Using Data Mining Techniques, CSE , Ambala City, India, An International Open Free Access, Peer Reviewed Research Journal .
4. VidyaPeetham Amrita Vishwa, Kasavanahalli, Carmelaram P.O. (2015):An Effective Performance Analysis of Machine Learning Techniques for Cardiovascular Disease, CSE , (Pg 23-32).
5. Mohammad A. M. Abushariah, Assal A. M. Alqudah, Omar Y. Adwan, Rana M. M. (2014): Automatic Heart Disease Diagnosis System Based on Artificial Neural Network (ANN) and Adaptive Neuro- Fuzzy Inference Systems(ANFIS),(Pg-Vol.7).
6. Salamay Mostafa A., Karam Omar H.,Khalifa M. Essam, El-Bialy Randa(2015):Feature Analysis of Coronary Artery Heart Disease Data Sets, CSE,British University in Egypt (BUE), Cairo, Egypt International Conference on Communication, Management and Information Technology,(Pg-459-468).
7. Pouriyeh Seyedamin, Vahid Sara, Sannino Giovanna, Arabnia Hamid, Gutierrez Juan, A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease, CSE, University of Georgia, Athens,
8. Babič František, Olejár Jaroslav(2017). Predictive and Descriptive Analysis for Heart Disease Diagnosis , Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, (pp. 155–163 DOI: 10.15439/2017F219 ISSN 2300-5963 ACSIS).
9. Methaila Aditya, Kansal Prince, Arya Himanshu(2014). EARLY HEART DISEASE PREDICTIONUSING DATA MINING TECHNIQUES ,CSE,Pankaj Kumar Netaji Subhas Institute of Technology,India,CCSEIT, DMDB, ICBB, MoWiN, AIAP – 2014 DOI : 10.5121/csit.2014.4807
10. Singh Poornima, Singh Sanjay, Pandi Gayatri S(2014). Effective heart disease prediction system using data mining techniques ,L. J. Institute of Engineering and Technology, Gujarat Technological University.
11. Khanna Divyansh, Sahu Rohan, Baths Veeky, Deshpande Bharat(2015). Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease, International Journal of Machine Learning and Computing.
12. Vahid Sara,Pouriyeh Seyedamin,Arabnia Hamid, Hamid Juan(2017). A comprehensive investigation, and comparison of Machine Learning Techniques in the domain of heart disease.