Ilkogretim Online - Elementary Education Online, 2020; Vol 19 (Issue 2): pp. 1817-1827
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2020.02.696765

Diagnosis And Prediction Of Heart Disease Using Machine Learning Techniques

J.Jeyaganesan[1] ; A.Sathiya[2] ; S.Keerthana[3] ; Aaradhyanidhi Aiyer[4]

[1]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[2]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[3]Asst professor Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

[4]Student, Department of AI&DS, Sri Sai Ram Institute of Technology, Chennai

ABSTRACT

The correct prediction of heart disease can prevent many life threats whereas incorrect prediction can be fatal at the same time. In my paper, I will be using different machine learning algorithms and data visualization techniques to predict heart disease using available dataset. The dataset consists of 14 main attributes used for performing analysis. I will be starting with data preprocessing step and will apply the machine learning algorithms on dataset of different sizes in order to study stability and accuracy and precision of each of them.

Keyword: Machine learning

Introduction

Heart disease is a range of conditions that can affect your heart. Nowadays, cardiovascular diseases are becoming the major cause of death worlwide with 17.9 million deaths yearly, as per the World unhealthy Health Organization reports. Various activities that lead to the risk of heart disease are high cholesterol, obesity, increase in triglycerides levels, hypertension, etc. There are certain signs which the American Heart Association lists like the persons having sleep issues, a certain increase and decrease in heart rate (irregular heartbeat), swollen legs, and in some cases weight gain occurring quite fast; it can be 1-2 kg daily.

Heart disease is very fatal and it should not be taken lightly. There are more chances of heart disease happening in males than in females.

Background

Heart disease affects millions of people, and it is the main cause of death in the world. Medical diagnosis should be efficient, reliable, and aided with computer techniques to reduce the effective cost for diagnostic tests. Data mining is a software technique that helps computers to build and classify various attributes. My research paper uses

different machine learning algorithms to predict and diagnose heart disease. In this, I have used different machine learning techniques and its methods data cleaning steps , evaluation and description of the dataset used in this research.

Machine Learning

Machine learning is a branch of Artificial Intelligence. Its primary focus is to design systems in such a way that it allows them to learn and make predictions based on the experience. It trains algorithms using a training dataset to create a model for better accuracy and prediction. The model uses the new input data to predict heart disease. By using machine learning, we can detect hidden patterns in the input dataset to build models. It gives accurate predictions for new datasets. Before using machine learning algorithm first we need to clean the dataset and fill the missing values. The model uses the new input data to predict heart disease and then tested for accuracy.

Machine learning is of 3 types which are classified as:

Supervised Learning

The model is trained on a dataset that is labelled. It has input data and its outcomes. Data elements are first classified and split in to train and test data. Training dataset used for training our model while testing dataset functions as new data to find the accuracy of the model. The two types under supervised learning are classification and regression.

Unsupervised Learning

Data elements used here for training are not classified or labelled. Aim is to find hidden patterns in the data. The model is trained to develop patterns. It can easily predict hidden patterns for any new input dataset or train data, but upon exploring data, it draws conclusion from datasets to describe hidden patterns. The clustering method is an example of an unsupervised learning technique.

Reinforcement Learning

It does not use any labelled dataset nor the results are associated with data, thus model learns based on the experience. In this technique, the model improves its performance based on its association with the environment and figures its faults and to get the right outcome through assessment and testing. Classification algorithms are commonly used supervised learning techniques to define probability of heart disease occurrence.

Classification

Machine Learning Techniques

The classification task is used for prediction of subsequent cases dependent on past information. Many data mining techniques as Naïve Bayes, neural network, decision

tree have been applied by researchers to have a precision diagnosis in heart disease. The accuracy provided by different techniques varies with number of attributes. This research provides diagnostic accuracy score for improvement of better health results. We have used WEKA tool in this research for pre-processing the dataset, which is in ARFF format (attribute-relation file format). Only 14 attributes of the 76 attributes have been considered for analysis to get accurate results. By comparing and analysing using different algorithms, heart disease can be predicted and cured early .

Approach Methodology

My research aims to make the prediction of having heart disease easy by using machine learning technique that is helpful in the medical field for clinicians and patients. To achieve my aim, I have discussed the use of various machine learning algorithms on the data set and dataset analysis is mentioned in my research paper. This paper also tells which attributes contribute more than the others for higher precision. This may spare the expense of different trials of a patient, as all the attributes may not contribute such a substantial amount to expect the outcome.

Description of the Dataset

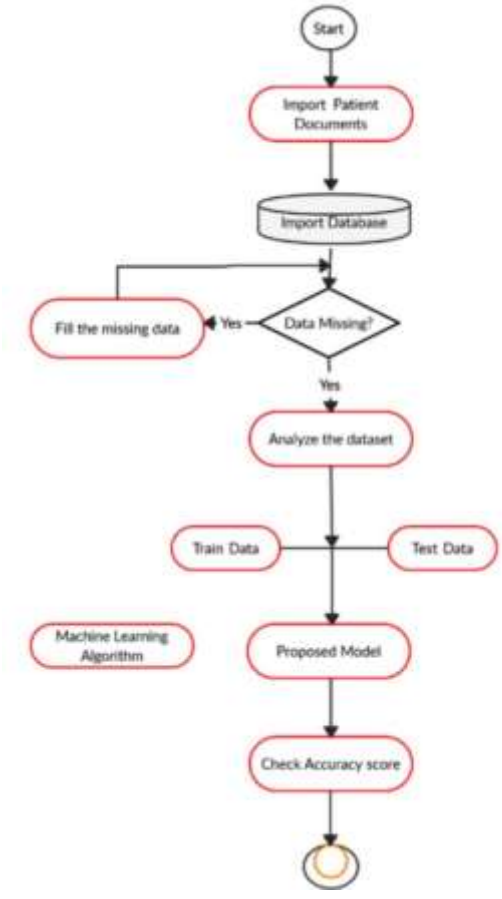
The dataset used for my research purpose was the Public Health Dataset which I have downloaded from kaggle . It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "output" or "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 which indicates no disease and 1 indicate disease. Now the attributes which are used in this research purpose are described as follows and for what they are used or resemble:
(i)Age—age of patient in years, sex—(1 represents male; 0 represents female).
(ii)Cp— represents chest pain.
(iii)Trestbps—represents resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if it is a little higher than the normal range it is risky, you should try to lower it).
(iv)Chol—represents serum cholesterol shows the amount of triglycerides present. Triglycerides is a lipid that is measured in the blood. It should be below than 170 mg/dL.
(v) Fbs—represents fasting blood sugar greater than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
(vi)Restecg—represents resting electro cardiographic results.
(vii)Thalach—maximum heart rate achieved. The maximum heart rate can be calculated by 220 minus your age.
(viii)Exang—represents exercise-induced angina (1 yes). Angina is a type of chest pain caused due to low blood flow to the heart. Angina can lead to coronary artery disease.
(ix)Oldpeak—represents ST depression induced by exercise relative to rest

(x)Slope—represents the slope of the peak exercise ST segment.
(xi)Ca—represents number of major vessels (0–3) coloured by fluoroscopy.
(xii)Thal—no explanation given in dataset, but probably represents thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
(xiii)Target (T)—no disease is represented by 0 and disease is represented by1, (angiographic disease status).

Data Pre-processing

The real-life information or data contains large numbers with missing and noisy data. These data are pre-processed to over-come such issues and make predictions vigorously. Figure explains the sequential chart of our proposed model.

Cleaning the collected data may contain missing values and may be noisy. To get an accurate and effective result, these data need to be cleaned in terms of noise and missing values are to be filled up. Transformation it changes the format of the data from one form to another to make it more comprehensible.



The dataset does not contain any null values. But many outliers are their which needs to be handled properly, and also the dataset is not properly distributed. Various plotting techniques were used for to check the distribution of the data, and outlier detection. All

these pre-processing techniques play an important role before we pass the data for classification or prediction purposes.

Checking the Distribution of the Data

The distribution of the data plays an important role when we need to the predict or classify a problem. This helps the model to find patterns in the dataset that leads to heart disease. 1 in graph represents people having heart disease and 0 represent people not having heart disease. Here 165 people suffer from heart disease and 138 people are without heart disease.

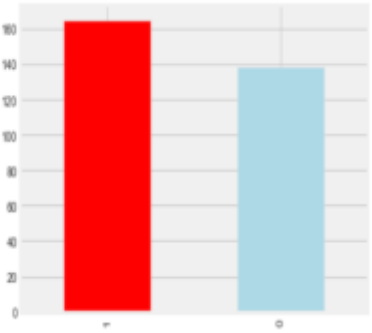


Fig 1:- Distribution of data

Analysis of Data

For Quantitative data:-

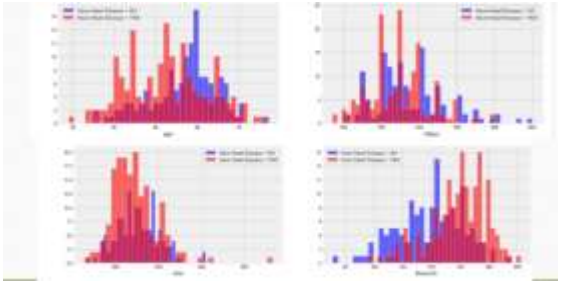


Fig 2:- Plots for numerical data

Observations that we can infer from the above plot are:

1. trestbps: represents resting blood pressure which above 130-140 is generally of concern
2. chol: if greater than 200 is of concern.

3. thalach: People with a higher value of over 140 are more likely to have heart disease.
4. The peak of exercise induced ST depression vs. rest looks at heart stress during exercise which tells that an unhealthy heart will stress more.

For Qualitative data:-

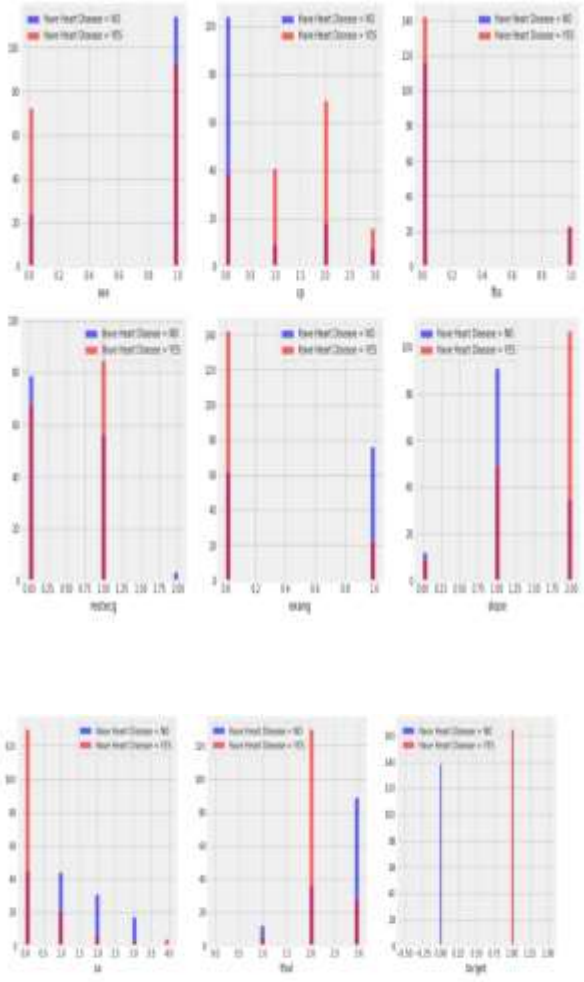


Fig3:- Plots for categorical data

Observations that we can infer from the above plot are:

1. cp {Chest pain}: People with cp 1, 2, 3 are more chances to have heart disease than people with cp 0.
2. restecg {resting EKG results}: People with a value of 1 (having an abnormal heart rate, which can range from mild symptoms to severe troubles) are more likely to have heart disease.
3. exang (exercise-induced angina): people with a value of 0 (No ==> angina induced by exercise) are more likely to have heart disease than people with a value of 1 (Yes ==> angina induced by exercise)

4. slope {the slope of the ST segment of peak exercise}: People with a slope value of 2 (Downslopins: indicates signs of an unhealthy heart) are more likely to have heart disease than people with a value of 2 slope is 0 (Upsloping: indicates best heart rate with exercise) or 1 (Flatsloping: indicates minimal change (typical healthy heart)).
5. ca [number of major vessels (0-3) stained by fluoroscopy]: the more movement of blood, people with ca equal to 0 are more likely to have heart disease.
6. thal {thalium stress result}: People with a value of 2 are more likely to have heart disease.

Correlation Matrix

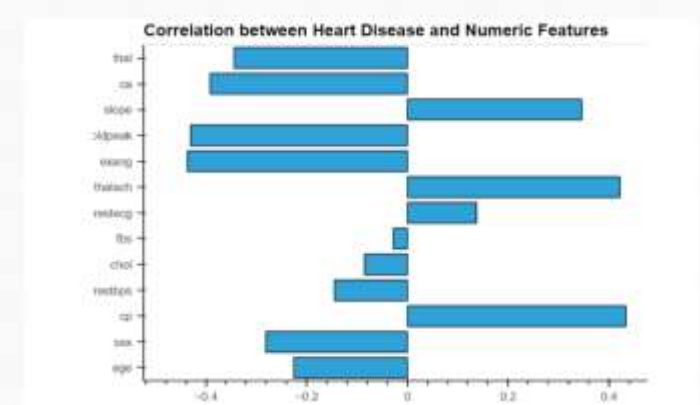


Fig 4:- Correlation Matrix

From the above fig we can observe that:-

1) fbs and chol are lowest correlated with the target variable.

2) All other variables have a significant correlation with the target variable.

Algorithm Used

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in thispaper are K nearest neighbours (KNN), Logistic Regression and Random Forest Classifiers which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This paper work includes examining the journals, published paper and the data of cardiovascular disease of the recent times. The methodology is a process which includes steps that transforms the input data into known data patterns for the knowledge of the users. The proposed methodology includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the pre-processing stages where we can explore the data. The main role of data pre-processing is to fill the missing values and clean the data. After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are K Nearest

neighbour(KNN), Logistic Regression, Random Forest Classifier. Finally, the proposed model is taken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in my model, an effective Heart Disease Prediction System (EHDPS) has been used using different classifiers. This model uses 13 health related parameters such as chest pain, blood pressure, blood pressure, age, cholesterol, fasting sugar, sex etc. for prediction.

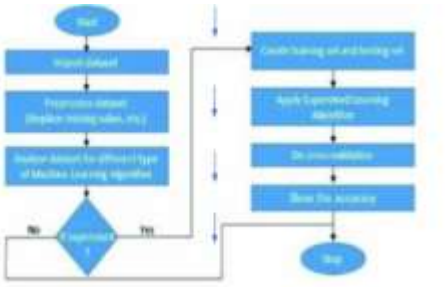


Fig 5:- Heart Disease prediction system flowchart

RESULTS & DISCUSSIONS

From the results I can infer that although most of the researchers have used different algorithms such as SVC, Decision tree for the detection of disease Random Forest Classifier provide a better result compared to them. The algorithms that I used are more accurate and save a lot of money i.e. it is cost efficient and provides fast results. Also, maximum accuracy was obtained by KNN and Logistic Regression which is almost more than 80% which is greater or almost equal to accuracies obtained from previous researches. So, to summarize that our accuracy may be improved due to the increased medical attributes that we used from the dataset we took.. The following 'figure 2', 'figure 3', 'figure 4','figure 5' shows a plot of the number of patient are been segregated and predicted by the classifier depending upon the age group, Resting Blood

Pressure, Sex, Chest Pain:

- Risk of Heart Attack
- No Risk of Heart Attack

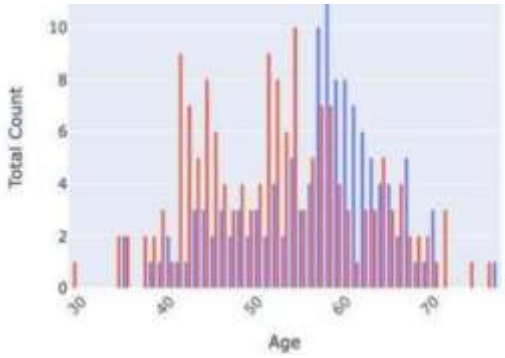


Fig 6:- Plot between age and total count

TABLE 1. Values Obtained for Confusion Matrix Using Different Algorithm

Algorithm	True Positive	False Positive	False Negative	True Negative
Logistic Regression	44	10	8	30
Naive Bayes	42	12	6	62
Random Forest	44	10	12	56
Decision Tree	50	4	8	60

TABLE 2. Analysis of Machine Learning Algorithm

Algorithm	Precision	Recall
Decision Tree	0.845	0.823
Logistic Regression	0.857	0.882
Random Forest	0.937	0.882
Naive Bayes	0.837	0.911

CONCLUSION

With the increasing number of deaths thanks to heart diseases, it's become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to seek out the foremost efficient ML algorithm for detection of heart diseases. This study gives a brief description about the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart

condition using the dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of more than 90% for prediction of heart disease. In future the work can be enhanced by developing an internet application supported the Random Forest algorithm also employing a large dataset as compared to the one utilized in this analysis which can provide better results and help health professionals in predicting the disease effectively and efficiently.

REFERENCES

[1] Sonam, A.M. “Predictions of Heart Condition Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June2016

vol-2

[3] Costas Sideris, Mohammad, Haik K, "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health

[4] Po Athi, Brad Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients Having Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pgno:1

[5] Dh, J K. Al, Mohamed Ibrahim, Mohammad Naeem "The Utilization of Machine Learning Approach for Medical Data Classification" in Annual Conference on New Trends in Information & Communication Technology Applications - march2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shou, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2,No. 3, June 2012

[7] Amu, J., Pad, S., Nandhini, R., Kavi, G., D, P., Venkata, V.S.K., "Recursive ant colony optimization routing in wireless mesh network",(2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Ala, B.P., Kavitha, A., Amu, J., "A novel encryption algorithm for end-to-end secured optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amu, J., In, P., B, B., Ananda, B., Ven, T., Prem, K., "An effective analysis on harmony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3),pp. 2035-2038.

[10]Amu, J., Kath, P., Reddy, L.S.S., Aa, A., "Assessment on authentication mechanisms in distributed system: A case study", (2017) Journalof Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11]Amu, J., Kode, C., Prem, K., Jai, S., Raja, D.,Ven, T., Hari, R., "Comprehensive analysis on information dissemination protocols in vehicular adhoc networks", 6 (2015) International Journal of Applied Engineering Research, 10 (3), pp.2058-2061.

[12]Amu, P., Reddy, L.S.S., Satyanarayana, K.V.V., "Effects, challenges.