Report

Major Project :

Heart disease prediction using machine learning algorithms

Team members:
1. Yash Tambat
2. Harshana Sundaram
3. Arya SS
4. Vishali Sudhakar
5. Maithre Sunitha
6. Dhanush Subramanian

Heart disease prediction using machine learning algorithms

Abstract. Day by day the cases of heart diseases are increasing at a rapid rate and it's very Important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. 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 bayes etc. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease It is implemented on the .pynb format.

1.Introduction

"Machine Learning is a way of Manipulating and extraction of implicit, previously unknown/known and

potential useful information about data" **[1].** Machine Learning is a very vast and diverse field and its scope

and implementation is increasing day by day. Machine learning Incorporates various classifiers of   Supervised, Unsupervised and Ensemble Learning which are used to predict and Find the Accuracy of the given dataset. We can use that knowledge in our project of HDPS as it will help a lot of people.

Cardiovascular diseases are very common these days, they describe a range of conditions that could affect your heart. World health organization estimates that 17.9 million global deaths from (Cardiovascular diseases) CVDs [2].

It is the primary reason of deaths in adults. Our project can help predict the people who are likely to diagnose with a heart disease by help of their medical history [6]. It recognizes who all are having any symptoms of heart disease such as chest pain or high blood pressure and can help in diagnosing disease with less medical tests and effective treatments, so that they can be cured accordingly.

This project focuses on mainly three data mining techniques namely: (1) Logistic regression, (2) KNN and (3) Random Forest Classifier. The accuracy of our project is 87.5% for which is better than previous system where only one data mining technique is used. So, using more data mining techniques increased the HDPS accuracy and efficiency. Logistic regression falls under the category of supervised learning. Only discrete values are used in logistic regression.

The objective of this project is to check whether the patient is likely to be diagnosed with any cardiovascular heart diseases based on their medical attributes such as gender, age, chest pain, fasting sugar level, etc. 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 a heart disease or not. To predict this, we use 14 medical attributes of a patient and classify him if the patient is likely to have a heart disease. These medical attributes are trained under three algorithms: Logistic regression, KNN and Random Forest Classifier. 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.

2. Related Work

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine

Learning algorithms has motivated this work. This paper contains a brief literature survey. An efficient

Cardiovascular disease prediction has been made by using various algorithms some of them include Logistic Regression, KNN, Random Forest Classifier Etc. It can be seen in Results that each algorithm has its strength to register the defined objectives [7].

The model incorporating IHDPS had the ability to calculate the decision boundary using the previous and new model of machine learning and deep learning. It facilitated the important and the most basic factors/knowledge such as family history connected with any heart disease. But the accuracy that was obtained in such IHDPS model was far more less than the new upcoming model such[2]as detecting coronary

heart disease using artificial neural network and other algorithms of machine and deep learning. The risk factors of coronary Heart disease or atherosclerosis is identified by McPherson et al**.,[8]** using the inbuilt implementation algorithm using uses some techniques of Neural Network and were just accurately able to predict whether the test patient is suffering from the given disease or not.

   Diagnosis and prediction of Heart Disease and Blood Pressure along with other attributes using the aid of neural networks was introduced by R. Subramanian et al.,**[24].** A deep Neural Network was Built

incorporating the given attributes related to the disease which were able to produce a output which was

carried out by the output perceptron and almost included 120 hidden layers which is the basic and most relevant technique of ensuring a accurate result of having heart disease if we use the model for Test Dataset. The supervised network has been advised for diagnosis of heart diseases [16]. When the testing of the model was done by a doctor using an unfamiliar data, the model used and trained from the previous learned data and predicted the result thereby calculating the accuracy of the given model.

3. Data Source

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and

in accordance with other medical conditions **[2].** Heart disease are the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle aged people are due to Cardiovascular diseases. We take a data source which is comprised of medical history of

304 different patient of different age groups. This dataset gives us the much-needed information i.e. the

medical attributes such as age, resting blood pressure, fasting sugar level etc. of the patient that helps us in detecting the patient that is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that helps us detecting if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having a heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patient prone to getting a heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in 'Table 1'.

Table 1. Various Attributes used are listed

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal
14.	Num(Disorder)	Heart Disease	Not Present /Present in the Four Major types.

4. Methodology

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this

paper are K nearest neighbors (KNN), Logistic Regression and Random Forest Classifiers which can be

helpful for practitioners or medical analysts for accurately diagnose Heart Disease. This

paperwork includes examining the journals, published paper and the data of cardiovascular

disease of the recent times. Methodology gives a framework for the proposed model **[13].** The

methodology is a process which includes steps that transform given data into recognized data

patterns for the knowledge of the users. The proposed methodology (Figure 1.) 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 preprocessing stage where we

explore the data. Data preprocessing deals with the missing values, cleaning of data and

normalization depending on algorithms used **[15].** After pre-processing of data, classifier is

used to classify the pre-processed data the classifier used in the proposed model are KNN,

Logistic Regression, Random Forest Classifier. Finally, the proposed model is undertaken,

where we evaluated our model on the basis of accuracy and performance using various

performance metrics. Here in this model, an effective **.**

4

(EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction **[17].**
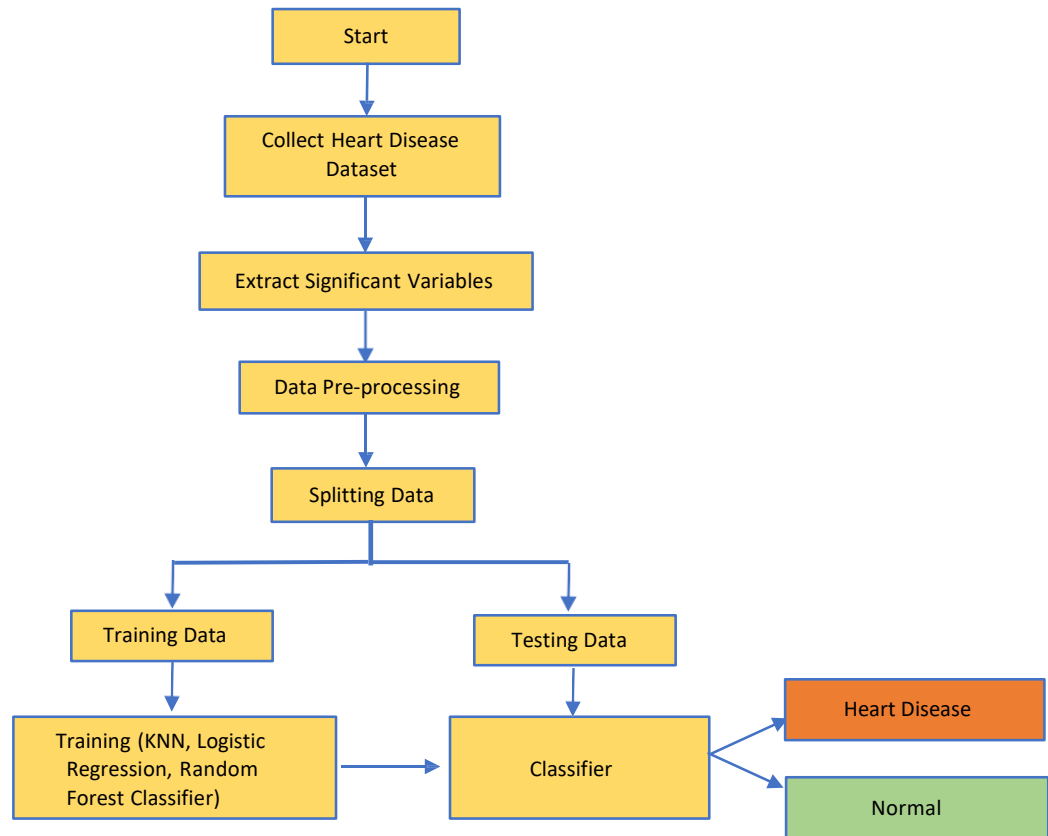

Figure 1. Proposed Model

5. Results & Discussions

From these results we can see that although most of the researchers are using different algorithms such as

SVC, Decision tree for the detection of patients diagnosed with Heart disease, KNN, Random Forest Classifier and Logistic regression yield a better result to out rule them **[23].** The algorithms that we used are more accurate, saves a lot of money i.e. it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by KNN and Logistic Regression are equal to 88.5% which is greater or almost equal to accuracies obtained from previous researches. So, we summarize

that our accuracy is improved due to the increased medical attributes that we used from the dataset we took. Our project also tells us that Logistic Regression and KNN outperforms Random Forest Classifier in the prediction of the patient diagnosed with a heart Disease. This proves that KNN and Logistic Regression are better in diagnosis of a heart disease. The following 'figure 2', 'figure 3', 'figure 4', 'figure 5' shows a plot of the number of patients that are been segregated and predicted by the classifier depending upon the age group, Resting Blood Pressure, Sex, Chest Pain:
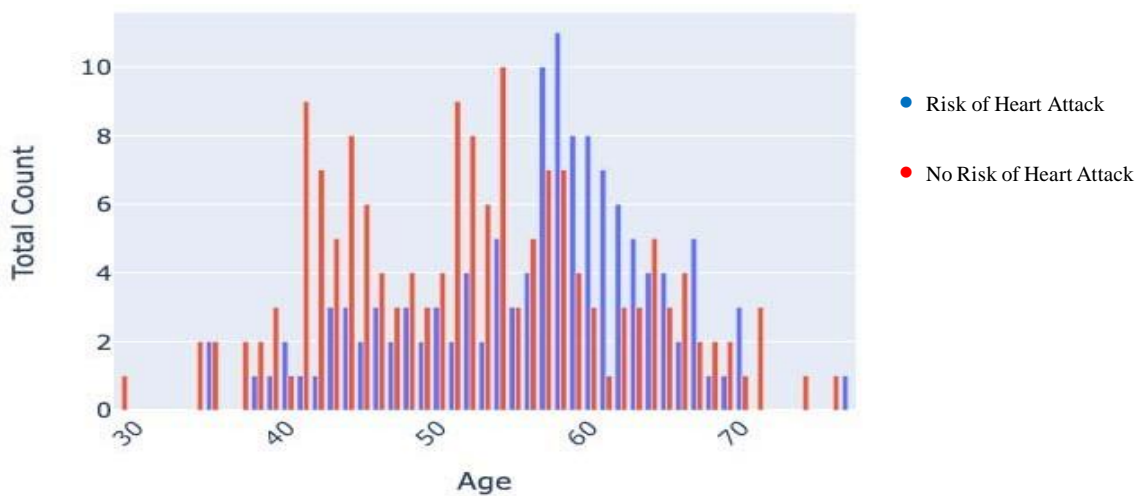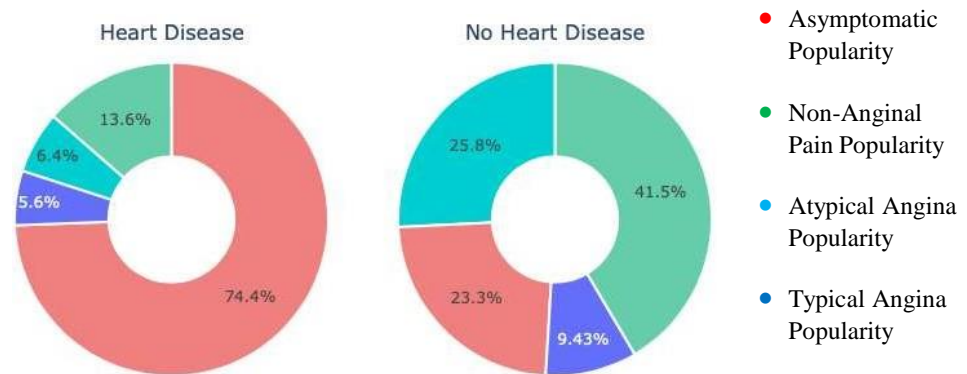

Figure 2. Shows the Risk of Heart Attack on the basis of their age.

Figure 3. Shows the Risk of Heart Attack on the basis of their Resting Blood Pressure.
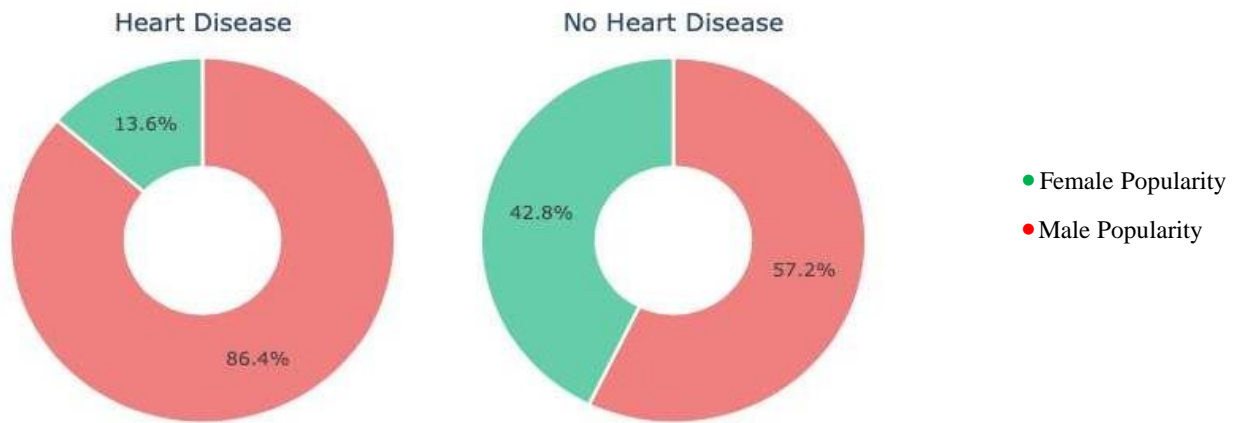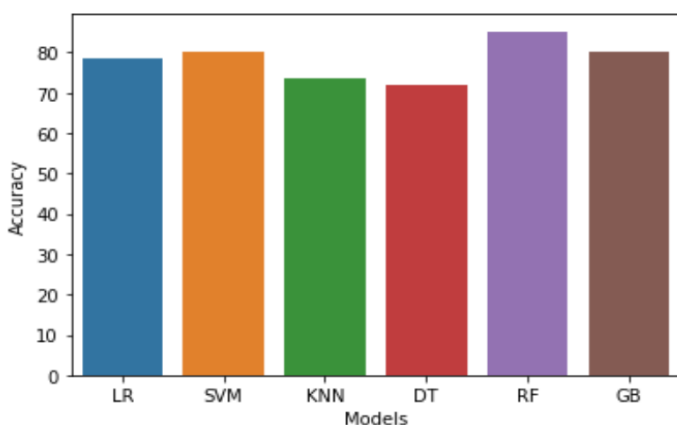
Figure 4. Shows the patients having or not having Heart Disease on the basis of Sex.

Figure 5. Shows the patients having or not having Heart Disease on the basis of type of Chest Pain.

Accuracy of Model :

`<AxesSubplot:xlabel='Models', ylabel='Accuracy'>`

1.Linear Regression : 78.68
2.Support Vector Machine : 80.32
3. KNN : 73.77
4. Decision Tree : 72.13
5.Random Forest : 85.24
6.Gredient Boosting : 80.32

6. Conclusion

A cardiovascular disease detection model has been developed using three ML classification modelling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and KNN [22]. The accuracy of our model is 87.5%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not [9]. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medical