International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

Heart Disease Prediction Using Machine Learning Techniques

Sai Bhavan Gubbala

[1]*Student, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology*

Abstract - *cardiovascular diseases are responsible for the majority of deaths globally. In most situations, the certainty of heart disease is known, but it may be anticipated early utilizing various machine learning methods. A proper diagnosis of heart patients can be made based on variables that are readily apparent. Age, diabetes, hypertension, cholesterol, obesity, and physical activity are a few of the contributing variables. In this study we are going to find an algorithm which predicts accurately. The dataset contains 14 attributes/features and 270 instances, out of which 150 have heart disease and 120 are considered to be normal. The data seems to be complete with no null values so we are going to use supervised machine learning models. To find out the best model we are going to compare five machine learning algorithms such as Random Forest classification, Support Vector Machine, AdaBoost Classifier, Logistic Regression and Decision Tree Classifier. The goal of this paper is to do comparative research on machine learning algorithms accuracy. This research is going to be developed further for best accuracy model prediction.*

Key Words: **Random Forest Classification, AdaBoost Classifier, Logistic Regression, Decision Tree Classifier, Support Vector Machine, Algorithm, Instances.**

1.INTRODUCTION

According to the World Health Organization, every year 15 million deaths occur worldwide due to heart disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques on the dataset available in the UCI repository, further evaluating the results using confusion matrix and cross validation. The author had used various machine learning techniques to predict the accuracy of test model such as Random Forest, Support Vector Machine, AdaBoost Classifier and we have clean dataset which can be readily used without cleaning or modifying the dataset.

2. LITERATURE OVERVIEW

There any many research works done on heart disease diagnosis by various factors. In this study, we are going to do a comparative analysis of different classification and regression algorithms and the results revealed that RF had obtained an higher accuracy which is the highest amongst the other algorithms. Authors had proposed a Random Forest Classification for Heart disease Prediction by integrating PCA and Cluster techniques. This study provided a significant contribution in computing strength scores with compelling predictors in heart disease prediction. Compared to five machine learning algorithms-Support Vector Machine, Decision Trees, Random Forest, AdaBoost Classifier and Logistic Regression-to predict heart disease. Among these algorithms, Random Forest gives the best accuracy, at 85.22% compared to other algorithms. The system evaluates all the parameters using the training and testing technique. The dataset is evaluated in python coding language. The coding language is then further processed in a Jupyter notebook which evaluates the process by a step-by-step manner. Different types of training and testing phases were implemented. In the end, the best pair of testing and training were selected and used in the process for achieving the highest accuracy.

3. DATASET INFORMATION

The dataset is collected from an online resource also known as the UCI repository. These days, the data is available every day and it is best to implement the model on a dataset which is available from a reliable resource. The dataset contains various attributes/features such as age, gender, fasting blood sugar, serum cholesterol, chest pain type, resting electrocardiographic results, exercise induced angina, ST depression, slope of peak exercise, resting blood pressure, number of major vessels, thalassemia and target. The dataset contains 270 instances with 14 attributes. The Table 1 gives us idea about how the dataset is implemented by numerical values. Figure 1 presents us a general overview of the dataset by indicating that 44% of patients are not diagnosed with heart disease and 56% of patients suffer cardiovascular disease.

Attributes	Description
Age	Age in years
Sex	Male-1, female-0
Chest Pain	Typical angina-0, Atypical angina-1, non-anginal pain-2, Asymptomatic-3
Resting Blood Pressure	94-200 mm in Hg
Cholesterol	126-524 mg/dl
Fasting Blood Sugar	True-1, False-0
Resting Electrocardiographic	Normal-0, ST-T wave abnoramility-1, Left ventricular hypertrophy-2
Maximum Heart	71-202 heart rate
Exercise Induced Angina	Yes-1, No-0
ST Depression	0-6.2 values
Slope of peak exercise ST segment	Upsloping-0, flat-1, downslopping-2
Number of major vessels	0-3 values
Thalassemia	0-normal,1-fixed defect,2-reversable defect
Heart disease (Target)	No-1, Yes-2

Table - 1: Dataset Attributes

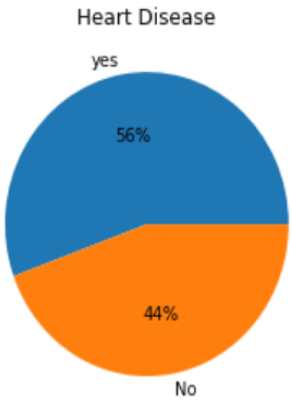

Fig - 1: Percentage of Heart Disease

3.1 DATASET PREPROCESSING

Data correlation, null value verification, loading python libraries, and partitioning the dataset into training and test portions are all steps in the preprocessing of the dataset. Preprocessing provides insight into how the characteristics of the data are influencing the data. Individual elements in the below correlation, testing, and training data are directly predictive of heart disease illness. People with heart disease are shown in Figure 2 along with their link to other relevant variables. The primary factor impacting heart disease has been found as the highest heart rate attained. Finally, Figure 3 provides us with a comprehensive connection of all features, which helps us identify the causes of heart disease. This understanding is exceptional because it distinguishes between big and small elements in regard to the goal value and is extremely helpful even when there is a link between them.

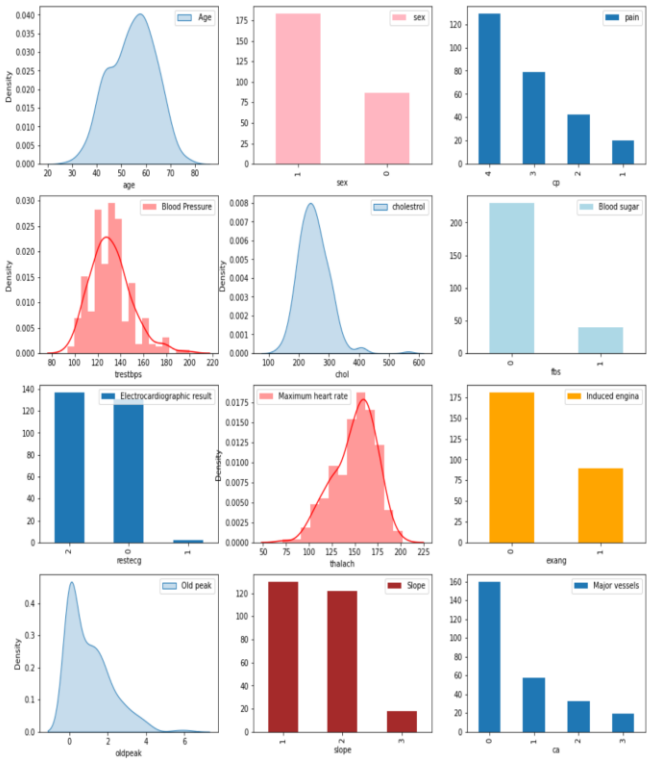

Fig -2: Attributes in respect to heart disease patients

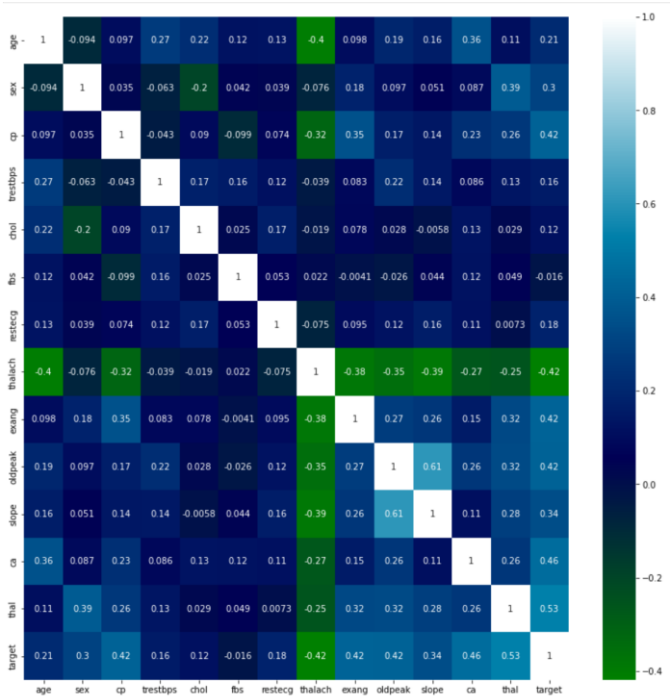

Fig – 3: Correlation of data

4. SPLITTING DATA INTO TESTING AND TRAINING

The training phase retrieves the characteristics from the data while the testing procedure determines the model behaviour for prediction. While the testing phase uses dependent factors, the training phase uses independent variables. There are many portions of the data. These are the steps of testing and training. 90% of the dataset is used for training, while 10% is used for testing. We consider the random state to be 1. To decide how the data will be split into train and test indices, we use the random state parameter. We can be certain that the method will always generate the same set of random integers if we set the random state to a constant value. The algorithms are already set up in a Jupyter notebook.

```
In [51]: X_train, X_test, y_train, y_test = train_test_split(features, target, test_size = 0.1, random_state = 0)
```

FIG - 4: Splitting data

5. METHODOLOGY

There are several sorts of algorithms, some of which differ from one another while others are similar; nonetheless, we now use all of these algorithms to compare their accuracy so that the one with the best accuracy may be chosen for improved predictability.

5.1 RANDOM FOREST CLASSIFIER

A large number of decision trees are built during the training phase of the random forests or random decision forests ensemble learning approach. It can used to solve for regression or classification problems. The Random Forest Classifier algorithm is used and its test accuracy was 85.22 percent.

5.2 DECISION TREE

Decision Tree is a structured classifier where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. It has two nodes such as Decision node and Leaf node. The test accuracy was 74.34 percent.

5.3 SUPPORT VECTOR MACHINE

Support Vector Machines implement nonlinear class boundaries using a linear model. The target classes are separated using support vectors. Before training a linear SVM model to classify the data in a higher-dimensional feature space to deal with a nonlinear condition, the model transforms the input using a mapping function. The test accuracy was 67.43 percent.

5.4 ADABOOST CLASSIFIER

An AdaBoost classifier is a meta-estimator that first fits a classifier on the original dataset, then fits additional copies of the classifier on the same dataset with the weights of instances that were incorrectly classified being changed so that subsequent classifiers concentrate more on challenging cases. The test accuracy was 78.59 percent.

5.5 LOGISTIC REGRESSION

Logistic Regression is also known as Analytical modelling technique. It is utilized to analyze datasets with one or more independent factors that affect the outcome. A random state of 0 was imported again for logistic regression. The training model was then fitted. The test's accuracy was 81.23 percent.

6. FLOWCHART OF DIAGRAM

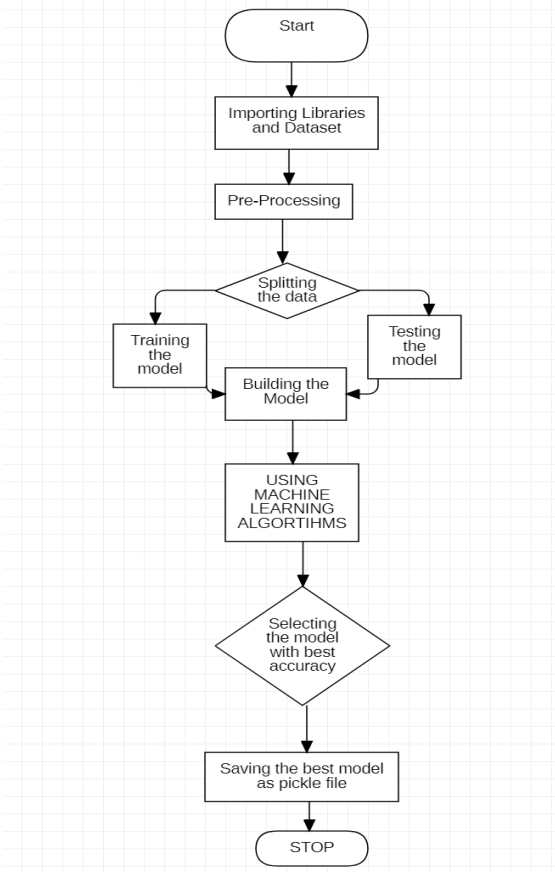

Fig -4: Flowchart of Research

The flowchart explains how the dataset is utilized while developing a prediction model. Understanding this study report requires understanding this flow chart.

7. RESULTS

According to the table, Random Forest has the best accuracy of all the classifiers, at 85.22 percent. Python was used to train the models, split the dataset into training and test data, and measure how accurate they were. Below is a comparison of the algorithms' results, along with a table listing their accuracy percentages.

Sl. No.	Algorithms	Accuracy
1.	Random Forest Classifier	85.22%
2.	Logistic Regression	81.23%
3.	AdaBoost Classifier	78.59%
4.	Decision Tree	74.34%
5.	Support Vector Machine	67.43%

Table -2: Performance evaluation of algorithms

```
RandomForestClassifier

classification_report :
              precision    recall  f1-score

           1       0.88      0.88      0.88
           2       0.80      0.80      0.80

    accuracy                           0.85
```

Fig – 5: Random Forest Accuracy

8. CONCLUSIONS

Therefore, we have finalized that Random Forest Classifier has attained the highest level of accuracy, as indicated by the numerical figures. Although Logistic Regression, AdaBoost Classifier, Decision Tree, and Support Vector Machine all had accuracy levels greater than 50%, they underperformed Random Forest. When compared to the other methods discussed, it can be reasoned that the random forest approach may also attain superior accuracy.

9.FUTURE SCOPE

Additionally, researchers may contrast this research's hierarchy levels to those from other datasets and draw interesting conclusions. The authors' intended study will aid in the creation of more accurate, reliable, and productive methods for predicting sickness, which will benefit not just the medical community but also numerous other groups and individuals throughout the globe.

REFERENCES

[1] Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J. A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE symposium on computers and communications (ISCC). IEEE. p. 204–207.

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(heart).

[3] "Prediction of Heart Disease using Machine Learning Algorithms" Krishnan J Santhana and S Geetha ICIICT |Year :2019| Conference Paper | Publisher: IEEE

[4] Bouali H, Akaichi J. Comparative study of different classification techniques: heart disease use case. In: 2014 13th international conference on machine learning and applications. IEEE. p. 482–86.

[5] World Health Organization. http://www.who.int/cardiovascular diseases/en. 2019.I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T.

Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[6] Chauhan R, Bajaj P, Choudhary K, Gigras Y. Framework to predict health diseases using attribute selection mechanism. In: 2015 2nd international conference on computing for sustainable global development (INDIACom). IEEE. p. 1880–84.

[7] "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" Senthil Kumar, Mohan Chandrasegar Thirumalai and Gautam Srivastva |Year :2019| Conference Paper | Publisher: IEEE.