International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 02 | Feb 2023 www.irjet.net p-ISSN: 2395-0072

Risk Of Heart Disease Prediction Using Machine Learning

Nikitha V P[1], Abhishek[2], Monika R[3], Monisha K[4], Vinodh Kumar S[5]

[2345]Students, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India
[1]Assoc. Professor, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India

---

Abstract – Today, the earth is revolving at same rate but people are evolving rapidly. And changing lifestyle of people is main cause for all impacts on human beings. Today, the death rate is 7.6 according to the survey and the main cause of most death is due to heart disease. The changing lifestyle of people have increased the effect on the normal functioning of heart.

This is caused in both men and women but it's more effective in men. In this generation people of all age group have been the predicted to have heart disease from newborn to old aged. The diagnosis isn't easy because death is occurring at first attack on heart. Predicting the risk of heart disease will have a best impact on curing the disease easily. This model helps in predicting of risk of heart disease in human at early stage and immediate precautions can be taken. This machine uses Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost and Artificial neural network. The random forest algorithm plays a vital role in prediction accuracy. The prediction of risk of heart disease can be done for all age groups to reduce the death rate.

Key Words: Machine learning, Logistic Regression, Naive Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, XGBoost and Artificial neural network.

1. INTRODUCTION

The World Health Organization estimates that heart disease accounts for 15 million deaths worldwide each year. Since a few years ago, the prevalence of cardiovascular disease has been rising quickly throughout the world. Numerous studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Even the silent killer of heart disease, which causes death without outward signs of illness, is addressed. In order to avoid complications in high-risk patients and make decisions about lifestyle changes, early detection of heart disease is crucial. Through the numerous cardiac characteristics of the patient, we have built and investigated models for risk of heart disease prediction in this project.

This model uses random forest algorithms by comparing 8 algorithm to predict the risk of heart disease. High accuracy is achieved in this model using Random Forest Algorithm. This Model takes various inputs and predicts the risk of heart disease. The output of the model is a target variable is binary value neither 0 or 1.

1.1 Motivation

As the heart disease being main cause for the deaths in the world and it's been realised that people from all age groups have been predicted to have heart disease. It's important to predict the risk of heart disease at early stage which helps in diagnosis. This model is developed to predict the risk which as a main strategy to reduce the death rate.

Also, this model can help in saving a life and alert people in taking care of themselves if they are at risk of heart disease.

1.2 Programming Language

Python is utilised for these projects. Because of its extensive libraries and ease of use for data analysis, manipulation, and artificial intelligence projects, application of AI and machine learning models, etc. In this study, Python frameworks are mentioned for the application of various methods for predicting risk of disease.

2. LITERATURE REVIEW

Numerous studies have been conducted on the diagnosis of heart disease using a variety of variables. In this study, we will compare various classification and regression algorithms. The results showed that RF had the highest accuracy among the algorithms, with a higher accuracy than the other algorithms. By combining PCA and Cluster methods, authors had suggested a Random Forest Classification for Heart Disease Prediction. This study made a substantial addition to the calculating of strength scores with convincing predictions in the prognosis of heart disease. Support Vector Machine, Decision Trees, Random Forest, AdaBoost Classifier, and Logistic Regression are five machine learning methods that have been compared to predict heart disease. When compared to other algorithms, Random Forest provides the highest accuracy, at 85.22%. The system uses the training and testing technique to evaluate all the parameters.

Python code is used to evaluate the dataset. A Jupyter notebook is used to process the coding language in more detail and evaluate the procedure step-by-step. Phases of testing and training of various kinds were used. In the end, the most accurate testing and training combinations were chosen and applied to the procedure.

3. DATASET INFORMATION

The dataset was compiled using a web-based tool called the UCI repository. These days, data is readily available every day, thus it is best to use a dataset that is obtainable from a trustworthy source while implementing the model. Age, gender, fasting blood sugar, serum cholesterol, type of chest pain, results of resting electrocardiograms, exercise-induced angina, ST depression, slope of peak exercise, resting blood pressure, number of major vessels, thalassemia, and target are just a few of the attributes and features included in the dataset. There are 270 instances in the collection with 14 attributes. Table 1 provides an overview of the dataset's use of numerical values. Figure 1 gives us a broad overview of the data by showing that 44% of patients do not have a heart disease diagnosis and 56% of patients suffer from this disease.

ATTRIBUTE	DESCRIPTION
Age	Age in years
Sex	Male-1, female-0
Chest pain type	Typical angina-0, Atypical angina-1, non-anginal pain-2, Asymptomatic-3
Resting blood pressure	94-200 mm in Hg
Maximum Heart	71-202 heart rate
Fasting Blood Sugar	True-1, False-0
Cholesterol	126-524 mg/dl
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-0, Yes-1
Resting electrocardiography	Normal-0, ST-Twave abnoramility-1, Left ventricular hypertrophy-2

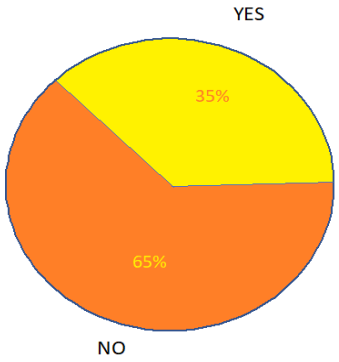


Fig-1: Percentage of heart disease

3.1 DATASET PREPROCESSING

The pre-processing of the dataset includes procedures such as data correlation, null value verification, loading python modules, and splitting the dataset into training and test halves. Pre-processing reveals how the traits are of the data are having an impact on the data. various components of the following testing, training, and correlation data are directly heart disease sickness prognostic. The main element influencing heart disease has been discovered to be the greatest heart rate ever. Qualities that aid in locating the root causes of cardiac disease.

This comprehension stands out because it makes a distinction between factors, both large and little, with relation to the goal value and is extremely helpful even when there is a link between them.

4. SPLITTING DATA INTO TESTING AND TRAINING

Data correlation, null value verification, loading python libraries, and partitioning the dataset into training and test portions are all steps in the pre-processing of the dataset. Pre-processing 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. The primary factor impacting heart disease has been found as the highest heart rate attained. Finally, 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.

5. METHODOLOGY

The model uses Random Forest algorithms that performs various process different from one another and while some are similar. All this algorithms are compared to choose the best algorithm to attain improved predictivity.

5.1 LOGISTIC REGRESSION

One of the most often used Machine Learning algorithms, within the category of Supervised Learning, is logistic regression. Using a predetermined set of independent factors, it is used to predict the categorical dependent variable. In a categorical dependent variable, the output is predicted via logistic regression. As a result, the result must be a discrete or categorical value. Rather than providing the exact values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc.

With the exception of how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems.

5.2 NAÏVE BAYES

The words Naive and Bayes, which make up the Nave Bayes algorithm, are as follows:

Because it presumes that the occurrence of one trait is unrelated to the occurrence of other features, it is referred to as naive. A red, spherical, sweet fruit, for instance, is recognised as an apple if the fruit is identified based on its colour, form, and flavour. So, without relying on one another, each characteristic helps to recognise it as an apple. Because it relies on the Bayes' Theorem concept, it is known as the Bayes principle.

5.3 SUPPORT VECTOR MACHINE

One of the most well-liked supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues.

The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.

SVM selects the extreme vectors and points that aid in the creation of the hyperplane. Support vectors, which are used to represent these extreme instances, form the basis for the SVM method. Take a look at the diagram below, where two distinct categories are identified using support vector Advantages of Misuse Attack Detection.

5.4 K-NEAREST NEIGHBORS

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K-NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

5.5 DECISION TREE

A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label.

By dividing the source set into subgroups based on an attribute value test, a tree can be "trained". It is known as recursive partitioning to repeat this operation on each derived subset. When the split no longer improves the predictions or when the subset at a node has the same value for the target variable, the recursion is finished.

5.6 RANDOM FOREST

Every decision tree has a significant variance, but when we mix them all in parallel, the variance is reduced since each decision tree is perfectly trained using that specific sample of data, and as a result, the output is dependent on numerous decision trees rather than just one. The majority voting classifier is used to determine the final output in a classification challenge. The final output in a regression problem is the mean of every output. This part is Aggression.

This method's fundamental principle is to integrate several decision trees to get the final result rather than depending solely on one decision tree.

Multiple decision trees serve as the fundamental learning models in Random Forest. We carry out random row sampling and further sampling will form datasets for model. This is called as bootstrap.

5.7 XGBOOST

XgBoost stands for Extreme Gradient Boosting. Gradient Boosted decision trees are implemented using XGBoost technology. Many Kaggle Competitions are dominated by XGBoost models.

Decision trees are generated sequentially in this approach. Weights are significant in XGBoost. Each independent variable is given a weight before being fed into the decision tree that forecasts outcomes. Variables that the tree incorrectly predicted are given more weight before being placed into the second decision tree. These distinct classifiers/predictors are then combined to produce a robust and accurate model. It can be used to solve problems including regression, classification, ranking, and custom prediction.

5.8 ARTIFICIAL NEURAL NETWORK

Artificial Neural Networks (ANN) are brain-inspired algorithms that are used to foresee problems and model complex patterns. The idea of biological neural networks in the human brain gave rise to the Artificial Neural Network (ANN), a deep learning technique. An effort to simulate how the human brain functions led to the creation of ANN. Although they are not exactly the same, the operations of ANN and biological neural networks are very similar. Only structured and numeric data are accepted by the ANN algorithm.

Unstructured and non-numeric data formats like image, text, and speech are accepted by convolutional neural networks (CNN) and recursive neural networks (RNN). The only subject of this article is artificial neural networks.

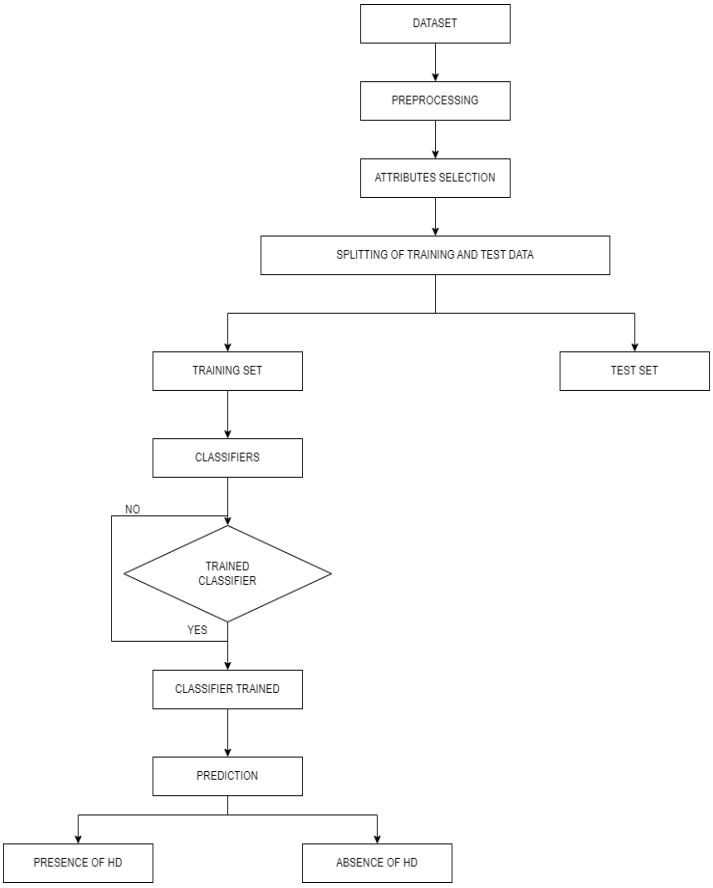
There are three layers in the artificial neural network architecture: the input layer, the hidden layer and the output layer.

6 SYSTEM DESIGN

The model is designed using random forest algorithms as mentioned above. All the algorithms are used for various processes like classification, structuring, analysis, prediction, decision making etc. The model takes 14 variables like age, sex, chest pain type, etc to predict the risk of heart disease as input to the algorithm. The algorithm classifies and makes a decision based on the classification.

The model gives an output as target value which contains binary value either 0 or 1. The prediction risk may indicate the presence of any heart disease related to vessels, structural problems and blood clots.

The below flow chart indicates the working of model:



7 CONCLUSION

The model predicts the risk of heart disease in humans which has become the main cause for the huge deaths occurring in the world. The system uses random forest algorithm to obtain a accuracy of 95%. The algorithm gives output as target value which is either 1 or 0 which indicates the presence if it's 1.

The model takes various attribute value to predict the output and each attribute forms a cause for the final output. Datasets are used as input which contains all the attributes values.

REFERENCES

[1] Preliminary Design of Estimation Heart disease by using machine learning ANN within one year DOI: 10.1109/rICT-ICeVT.2013.6741541

[2] Prediction of Heart Disease Using Learning Vector Quantization Algorithm DOI: 10.1109/CSIBIG.2014.7056973

[3] Using the Extreme Learning Machine (ELM) technique for heart disease diagnosis DOI: 10.1109/IHTC.2015.7238043

[4] Study of Machine Learning Algorithms for Special Disease Prediction using Principal of component analysis DOI: 10.1109/ICGTSPICC.2016.7955260

[5] Coupling if fast Fourier transformation using machine learning ensemble model Support recommendation for Heart disease patients in a telehealth environment DOI: 10.1109/ACCESS.2017.2706318

[6] Prediction of Heart Disease Using Machine Learning DOI: 10.1109/ICECA.2018.8474922

[7] Prediction of Heart Disease Using Machine Learning Algorithms DOI: 10.1109/ICIICT1.2019.8741465

[8] Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare DOI: 10.1109/ACCESS.2020.3001149