Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 11, November-2020

Heart Disease Prediction using Machine Learning Techniques

Pooja Anbuselvan
Student
Bangalore Institute of Technology
Bengaluru, Karnataka, India.

***Abstract*—Heart diseases also known as cardio vascular diseases encompass a wide range of conditions that affect the heart. These vary from blood vessel diseases, heart rhythm problems to heart defects that one is born with. It is the primary cause for death worldwide over the past few decades. It is the need of the hour to obtain accurate and reliable approach to achieve early diagnosis of the disease by automating the task and hence realize efficient management of it. Data Science plays an important in processing large amounts of data in the field of medical sciences. Researchers utilize several Data Mining and Machine Learning Techniques to analyze large sets of data and aid in the right prediction of heart diseases. This paper analyzes the supervised learning models of Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest and the ensemble technique of XGBoost to present a comparative study for the most efficient algorithm. It is found that Random Forest provides most accuracy with 86.89% in comparison to other algorithms.**

Keywords*—Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, XGBoost, Heart Disease Prediction.

I. INTRODUCTION

Cardio-Vascular diseases are the primary cause of death worldwide over the past decade. According to the World Health Organization it is estimated that over 17.9 million death occur each year because of cardiovascular diseases and out of these deaths 80% is attributed to coronary artery disease and cerebral stroke [1]. Many habitual factors such as personal and professional habits and genetic predisposition accounts for heart disease. Various risk factors such as smoking, overuse of alcohol and caffeine, stress, and physical inactivity along with other physiological factors like obesity, hypertension, high blood cholesterol, and pre-existing heart conditions are often deciding factors for heart disease. The efficient, accurate and early medical diagnosis of heart disease plays a pivotal role in taking preventive measures to avoid the complications that arise due to such diseases.

The major challenge faced in the world of medical sciences today is the provision of quality service and efficient and accurate prediction. The later problem can be solved by automation with the help of Data Mining and Machine Learning. Data mining is defined as a process used to extract usable data from a large set of raw data. It implies analyzing patterns in large batches of data by making use of various software. It also involves effective data collection and warehousing coupled with computer processing. Machine Learning (ML) which is subfield of data mining that deals with large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases.

Various Machine Learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest and the ensemble technique of XGBoost are compared to find the most accurate model. Here the heart disease dataset from the UCI repository is used. In this research a discussion and comparison of the existing classification techniques is made. The paper also mentions scope of future research and different advancement possibilities.

II. RELATED WORK

- Gomathi et al. used Naive Bayes and decision tree data mining techniques for predicting different types of diseases. They mainly concentrated on prediction of heart diseases, diabetics and breast cancer. The results were derived from the confusion metrics.
- Miranda et al. proposed Naive Bayes classifier approach for the prediction of the cardiovascular diseases. The authors have considered few important risk factors for deciding the heart disease.
- Avinash Golande and et. al.;studies various different ML algorithms that can be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.
- Fahd Saleh Alotaibi has designed a ML model comparing five different algorithms. Rapid Miner tool was used which resulted in higher accuracy compared to Matlab and Weka tool. In this research the accuracy of Decision Tree, Logistic Regression, Random forest, Naive Bayes and SVM classification algorithms were compared. Decision tree algorithm had the highest accuracy.
- Theresa Princy. R, et al, executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K- Nearest Neighbour),

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 11, November-2020

Decision tree, Neural network and accuracy of the classifiers was analyzed for different number of attributes.

III. DATA RESOURCE

The Cleveland heart dataset from the UCI machine learning repository has been used for the experiments. The dataset consists of 14 attributes and 303 instances. There are 8 categorical attributes and 6 numeric attributes. The description of the dataset is shown in the table.

Attribute	Description	Range
Age	Age of person in years	29-79
Sex	Gender of person (1-M 0-F)	0,1
Cp	Chest pain type	1,2,3,4
Trestbps	Resting blood pressure in mm Hg	94-200
Chol	Serum cholesterol in mg/dl	126-564
Fbs	Fasting blood sugar in mg/dl	0,1
Restecg	Resting Electrocardiographic results	0,1,2
Thalach	Maximum heart rate achieved	71-202
Exang	Exercise Induced Angina	0,1
OldPeak	ST depression induced by exercise relative to rest	1-3
Slope	Slope of the Peak Exercise ST segment	1,2,3
Ca	Number of major vessels colored by fluoroscopy	0-3
Thal	3 – Normal, 6 – Fixed Defect, 7 – Reversible Defect	3,6,7
Result	Class Attribute	0,1

Patients from age 29 to 79 have been selected in this dataset. Male patients are denoted by a gender value 1 and female patients are denoted by gender value 0. Four types of chest pain can be considered as indicative of heart disease. Type 1 angina is caused by reduced blood flow to the heart muscles because of narrowed coronary arteries. Type 1 Angina is a chest pain that occurs during mental or emotional stress. Non-angina chest pain may be caused due to various reasons and may not often be due to actual heart disease. The fourth type, Asymptomatic, may not be a symptom of heart disease. The next attribute *trestbps* is the reading of the resting blood pressure. *Chol* is the cholesterol level. *Fbs* is the fasting blood sugar level; the value is assigned as 1 if the fasting blood sugar is below 120 mg/dl and 0 if it is above. *Restecg* is the resting electrocardiographic result, *thalach* is the maximum heart rate, *exang* is the exercise induced angina which is recorded as 1 if there is pain and 0 if there is no pain, *oldpeak* is the ST depression induced by exercise, *slope* is the slope of the peak exercise ST segment, *ca* is the number of major vessels colored by fluoroscopy, *thal* is the duration of the exercise test in minutes, and *num* is the class attribute. The class attribute has a value of 0 for normal and 1 for patients diagnosed with heart disease.

IV. APPROACH METHODOLOGY

A. Classification Algorithms

Classification is a supervised learning procedure that is used for predicting the outcome from existing data. This paper proposes an approach for the diagnosis of heart disease using classification algorithms. The dataset has been divided into a training set and a test set in ratio of eighty and twenty, and individual classifiers are trained using the training dataset. The efficiency of the classifiers is tested with the test dataset. The working of the individual classifiers is explained in the proceeding section.

1. Logistic Regression

   Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes 0 for failure and 1 for success

2. Naïve Bayes

   Naïve Bayes classifier is a supervised algorithm. It is a simple classification technique using Bayes theorem. It assumes independence among attributes. Bayes theorem is a mathematical concept that is used to obtain the probability. The predictors are neither related to each other nor have correlation to one another. All the attributes independently contribute to the probability to maximize it. Many complex real-world situations use Naive Bayes classifiers

   $$P(X/Y) = P(Y/X)\times P(X)/P(Y),$$

   $P(X/Y)$ is the posterior probability, $P(X)$ is the class prior probability, $P(Y)$ is the predictor prior probability, $P(Y/X)$ is the likelihood, probability of predictor.

3. Support Vector Machine

   An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

4. K-Nearest Neighbour

   The K-Nearest Neighbour algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbour. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbours is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them. K-NN algorithm is simple to carry out without creating a model or making other assumptions. This algorithm is versatile and is used for classification, regression, and search. Even though K-NN is the simplest algorithm, noisy and irrelevant features affect its accuracy.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 11, November-2020

5. Decision Tree

Decision tree is a classification algorithm that works on categorical as well as numerical data. Decision tree is used for creating tree-like structures It is easy to implement and analyze the data in tree-shaped graph.

This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gain or minimum entropy:

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(Y,X) = E(Y) - E(Y|X)$$

The results obtained are easier to read and interpret. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph. However, the data may be over classified and only one attribute is tested at a time for decision-making.

6. Random Forest

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more the number of trees higher is the accuracy. It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

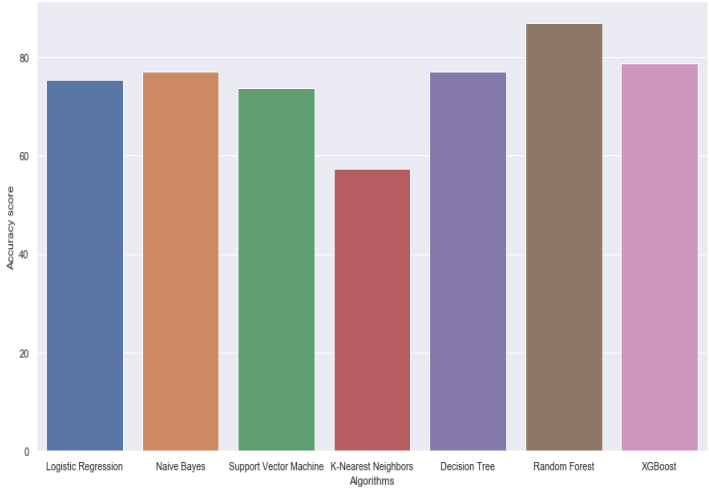
7. XGBoost

XGBoost is an optimized distributed gradient model designed to be highly efficient, flexible and portable. It is a decision-tree based ensemble Machine Learning algorithm that uses gradient boosting framework. It provides an optimized gradient boosting algorithm through parallel processing, tree pruning, handling missing values and regularization to avoid overfitting or bias.

V. RESULT AND ANALYSIS

The aim of this research is to analyze the performance of various classification algorithms and in doing so find the most accurate algorithm for predicting whether a patient would develop and heart disease or not. This research was done using techniques of Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Forest, XGBoost on the UCI dataset. Dataset was split into training and test data and models were trained and the accuracy was noted using Python. A comparison of the performance of the algorithms are depicted below and their accuracy scores are presented in the table.

Algorithm	Accuracy
Logistic Regression	75.41%
Naïve Bayes	77.05%
Support Vector Machine	73.77%
K-Nearest Neighbor	57.83%
Decision Tree	77.05%
Random Forest	86.89%
XGBoost	78.69%



VI. CONCLUSION

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is the goal of this research. The data were pre-processed and then used in the model. Random Forest with 86.89% and XGBoost with 78.69% are the most efficient algorithms. However, K-Nearest Neighbor performed with the worst accuracy with 57.83%. We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules and other ensemble techniques. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.

REFERENCES.

[1] Seckeler MD, Hoke TR. The worldwide epidemiology of acute rheumatic fever and rheumatic heart disease. Clin Epidemiol. 2011;3:67.

[2] Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE. 2017;12(4):e0174944.

[3] Ramalingam VV, Dandapath A, Raja MK. Heart disease prediction using machine learning techniques: a survey. Int J Eng Technol. 2018;7(2.8):684–7.

[4] 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.

[5] 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.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 11, November-2020

[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] T.Nagamani, S.Logeswari, B.Gomathy, Heart Disease Prediction using Data Mining with Mapreduce Algorithm, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.