International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Random Forest Based Heart Disease Prediction

Adeen[1], Preeti Sondhi[2]

[1]M. Tech Scholar, Universal Group of Institutions, Lalru, Punjab, India
adeenmir1995.am[at]gmail.com

[2]Assistant Professor, Universal Group of Institutions, Lalru, Punjab, India
preetisondhi4[at]gmail.com

Abstract: ***The key explanation for a large number of deaths in the world over the last few decades is heart-related diseases or cardiovascular diseases (CVDs), which have emerged as the most life-threatening disease not just in India, but in the world as a whole. So, in order to identify such diseases in time for proper care, there is a need for a reliable, precise and feasible method. Algorithms and methods of machine learning have been applied to large data sets in the field of medicine for data processing. Several data mining and machine learning techniques are used by researchers to analyse vast data sets and assist in the accurate prediction of heart diseases. This paper analyses the Naïve Bayes, Help Vector Machine, Random Forest, supervised learning models to present a comparative analysis for the most effective algorithm. Random Forest has been found to have 95.08% more precision compared to other algorithms.***

Keywords: Heart diseases, Naïve Bayes, Support Vector Machine, Random Forest, Heart Disease Prediction

1. Introduction

The heart is an essential organ in all living beings, and plays a crucial role in pumping blood into the circulatory system's blood vessels to the rest of the organs. Heart Disorders characterise a number of conditions that affect the heart and are the world's leading cause of death. Together, heart disease and stroke are today's most serious and expensive health problems facing the country. The heart has four valves that open and close to regulate blood flow into your heart: the aortic, mitral, pulmonary and tricuspid valves. A variety of conditions leading to narrowing, leaking (regurgitation or insufficiency) or improper closing can harm valves[1].While heart disease is often known as a "man's disease," in the United States, around the same number of women and men die each year of heart disease. About 1 in 4 women will die within the first year of a heart attack, compared with 1 in 5 men. Some diseases and lifestyle habits raise [2][3]the risk of a person developing heart disease, including diabetes, obesity and obesity, poor nutrition, physical inactivity, and heavy use of alcohol.

One of the major organs responsible for the functioning of blood in our human body and for all parts of the body is heart, one of the most common disease in India is heart attack, if the heart stops functioning then the complete blood circulation system in our body stops, which can also cause death, leading to serious health condition. The following are the forms of heart disease generally found in the world in which a class of diseases affecting the heart or blood vessels is Cardio Vascular Disease. CVD involves coronary artery disease (CAD), also known as heart attack, such as angina and myocardial infarction. Coronary heart disease is the other kind of heart disease[4] of which plaque build-up is the usual cause. This causes the coronary arteries to widen, restricting the heart's blood flow.

The following are the symptoms of heart attack

1) Tiredness: The person feels like sweating throughout and looks very tired.
2) Lack of Oxygen: level of the oxygen drops causes dizziness and out of balance.
3) Bradycardia: in this patient will have a slower heart rate of over 50-60 bpm.
4) Hypertension: In this the patient heart rate varies from 100-200 bpm and over some times.
5) Chest pain: The chest pain happens because of the blockage in the vessel of coronary part of the body.
6) Pain in Arm: The pain usually starts in the chest part and slowly it creeps to the arm.

Attributes we have taken for the analysis

```
age:        age
sex:        1: male, 0: female
cp:         chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic
trestbps:          resting blood pressure
chol:       serum cholestoral in mg/dl
fbs:        fasting blood sugar > 120 mg/dl
restecg:           resting electrocardiographic results (values 0,1,2)
thalach:           maximum heart rate achieved
exang:      exercise induced angina
oldpeak:           oldpeak = ST depression induced by exercise relative to rest
slope:      the slope of the peak exercise ST segment
ca:         number of major vessels (0-3) colored by flourosopy
thal:       thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
```

Figure 1: Dataset Attributes

Licensed Under Creative Commons Attribution CC BY

2. Techniques and Algorithms

Logistic Regression

Logistic Regression: Logistic regression is a binary regression model used for categorical data and logistic regression is used also in the case of continuous data in statistics to model the likelihood of a certain class or event such as good or sick, pass or fail, etc. A supervised learning classification algorithm used to predict the likelihood of a target variable is logistic regression [5]. The essence of the goal or dependent variable is dichotomous, meaning that only two possible grades will be 0 for failure and 1 for achievement. To model the data and infer the outcomes in the form of a curve, we use a sigmoid function.

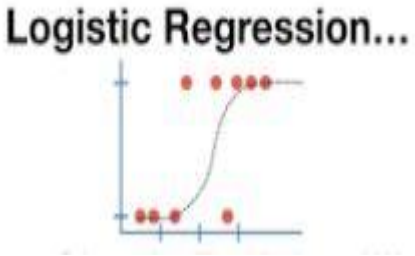

Figure 2: Logistic regression

Naive Bayes

Naïve Bayes: A supervised algorithm is a Naïve Bayes classifier. It is a basic technique for classification using the Bayes theorem. It implies the attributes are independent. The theorem of Bayes is a mathematical notion used to obtain probability. The predictors are neither related nor associated to each other. Independently, all the attributes contribute to the likelihood of optimising it. Many complex real-world scenarios use classifiers from Naive Bayes[6][7][8]. This algorithm is one of the simplest and better classification techniques for properly classifying objects, based on the Bayes theorem that this technique uses independent attributes or variables for data calculation and conclusions that also use the conditional probability function.

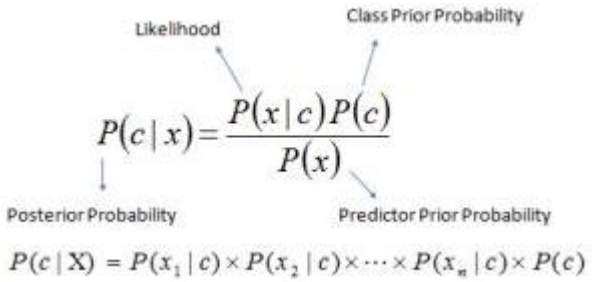

Figure 3: Naïve Bayes

Support Vector Machine

A support vector machine is a common algorithm for supervised machine learning used as a classifier and predictor; it uses a hyper plane used to distinguish classes. The training data points are represented by an SVM model as points in the feature space, mapped in such a way that points belonging to different classes are separated as broadly as possible by a margin. In the same space, the test data points are then mapped and categorised depending on which side of the margin[9] they fall. In multidimensional space, an SVM model is essentially a representation of different classes in a hyper plane. To minimise the error, the hyper plane will be generated by SVM in an iterative manner. SVM aims to break the datasets into groups in order to find the optimal marginal hyper plane (MMH).

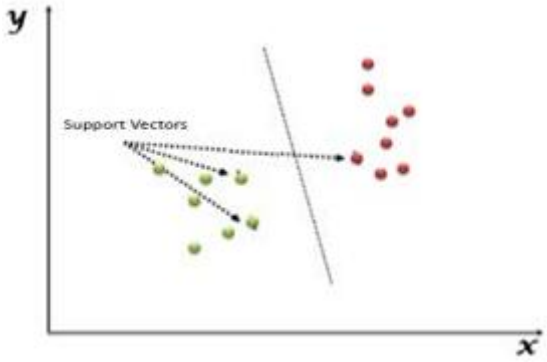

Figure 4: Support vector machine

Decision Trees

Decision tree is a supervised learning algorithm, it is used for classification tasks in which the categorical variables are primarily used. In this process, the entropy is calculated for each and every attribute and later the data set is split based on minimum entropy or maximum knowledge gain. The main population is divided into two or more similar sets based on more predictor values [7][9]. A classification algorithm that operates on both categorical and numerical data is a decision tree. The decision tree is used to construct tree-like structures In a tree-shaped graph, the data can be easily applied and analysed.

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

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Figure 5: Entropy and gain equation

Random Forest

This technique is both a classification and regression-oriented technique, as the name suggests that before finalising a performance criterion, it is a mixture or set of multiple decision trees. The main purpose of this technique is to conclude that more trees can cover the correct decision. It uses a voting method for classification to determine the output and it takes mean of all the outputs of each of the decision trees for regression, it works well with broad and high dimensional results. It is also known under Learning ensemble. The Random Forest Algorithm is an algorithmic method for supervised classification. Several trees build a forest inside this algorithm. In a random forest, each individual tree sets out a class expectation and the class with the most votes becomes the forecast of a model. The greater the number of trees in the random forest classifier, the greater the precision. It is used for the task of classification and regression, but can do well with the task of classification, and can solve missing values.

Licensed Under Creative Commons Attribution CC BY

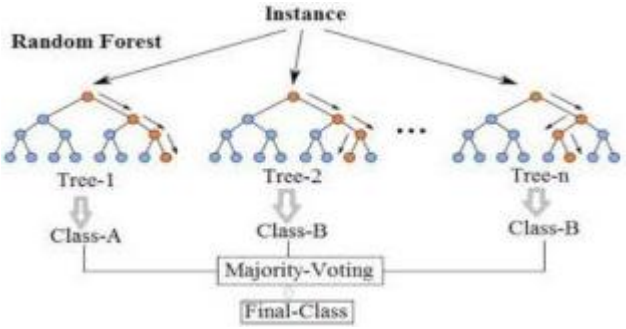

Figure 6: Random Forest

3. Result and Discussion

The purpose of this analysis is to evaluate the performance of different classification algorithms and thus find the most effective algorithm to predict whether or not a patient may develop heart disease. This work was carried out on the dataset using Naïve Bayes, Support Vector Machine and Random Forest techniques. Dataset has been split into training and test data and models have been trained and using Python the accuracy has been noted. A comparison of the performance of the algorithms is shown below and the table presents their accuracy ratings, recall, accuracy, F1 score.

Table 1: Naïve Bayes, SVM, Random Forest Comparison Table

Algorithm	F1 Score	Recall	Precision	Accuracy
Naïve Bayes	0.87	0.91	0.84	85.25%
Support vector machine	0.85	0.88	0.81	81.97%
Random Forest	0.95	0.91	1	95.08%

```
scores = [score_nb,score_svm,score_rf]
algorithms = ["Naive Bayes","Support Vector Machine","Random Forest"]
for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using Random Forest is: 95.08 %
```

Figure 7: Accuracy score

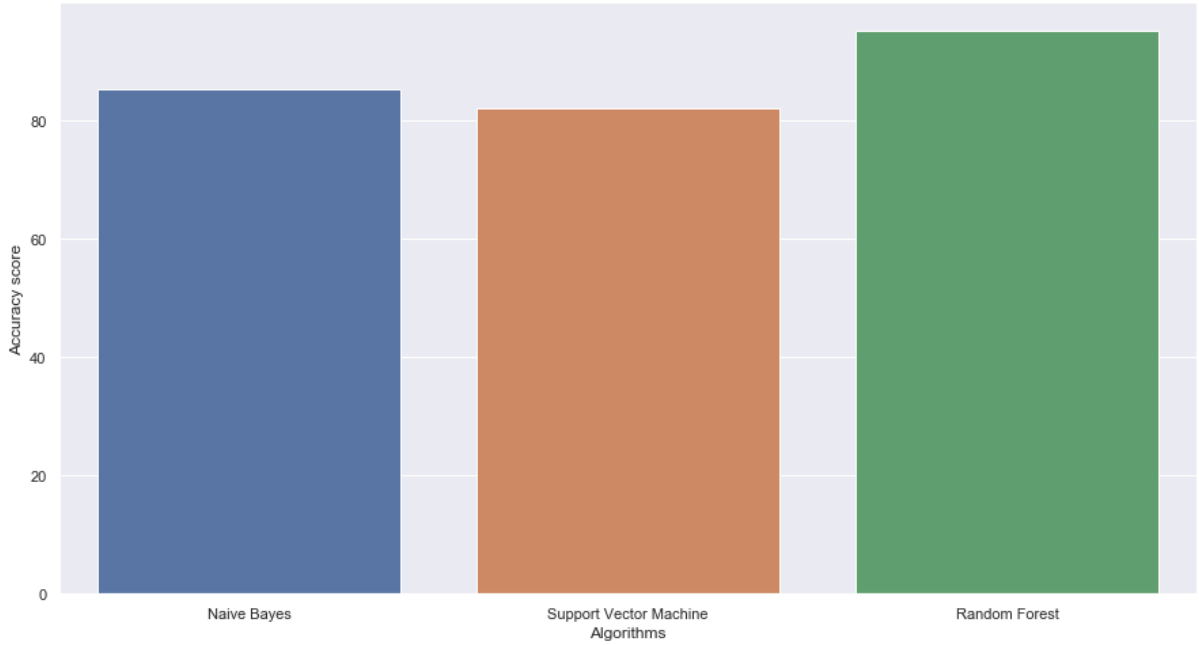

Figure 8: Comparison graph

4. Conclusion and Future Work

On the basis of the above analysis, it is clear that most machine learning algorithms perform well in the prediction and diagnosis of cardio vascular or heart diseases, some algorithms may be weak in their outcomes in terms of efficiency and accuracy tests, Random Forest algorithm works well on over fitting data, while algorithms such as support vector machine and Naive Bayes algorithm work well on overfitting data. In future we will expand the use of these algorithms to high-dimensional data and use neural network for better accuracy.

References

[1] V Ramalingam, Dandapath, Karthik Raja, Heart Disease Prediction Using Machine Learning Techniques: a survey, International Journal of Engineering and Technology, 2018.

Licensed Under Creative Commons Attribution CC BY

[2] Nandhni, Debnath, Pushkar, Heart disease prediction using machine learning, International Jour of Engineering Research and Development, 2018.
[3] Tamara saad Mohammad, Heart disease prediction using weka, Baghdad College of economic sciences university, issue 58.
[4] M A Jabbar, BL Deekshithulu, Heart Disease prediction using Genetic algorithm based trained recurrent fuzzy neural networks ,International conference on theory and application of soft computing with words and perception, 2017.
[5] M muthuvel, Analysis of Heart disease prediction using various machine learning techniques, international conference on artificial intelligence smart grid, smart applications,2019.
[6] AbhisheikRairikar, vedantkulkarni, Heart disease prediction using data mining Techniques, IEEE conference on intelligent computing & control,2017.
[7] Avinash G, Heart disease prediction using effective machine learning Techniques, International jour of recent technology and engineering, 2019.
[8] Reddy Prasad, Anjali, Deepa, Heart disease prediction using logistic regression using machine learning, IJEAT,2019.
[9] Dinesh Kumar, prediction of cardiovascular disease using machine learning algorithms, IEEE conference 2018.
[10] 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.

Licensed Under Creative Commons Attribution CC BY