Comparative Analysis of Machine Learning Techniques in Heart Disease Prediction by R Language

Avni Sharma[1] Deeksha Tyagi[2] Dr. Tarun Kumar Gupta[3]
[1,2]Student [3]Professor
[1,2,3]Department of Computer Science and Engineering
[1,2,3]Radha Govind Engineering College, Meerut, Uttar Pradesh, India

Abstract— Heart disease is the leading cause, which has accounted serious death rate, worldwide. A Big amount of data is present in medical industry, which has been continuously used by researchers to produce new scientific techniques to reduce number of deaths from heart diseases. There is a need of an efficient scientific technique, in order to simplify this alarming problem. This paper presents a Heart disease prediction model that can help medicinal experts in anticipating Heart sickness status based on clinical information of patients. This study compares different machine learning algorithms seeking better performances in heart disease prediction using R language. The algorithms which are used i.e... Logistic Regression Model, Random Forest Tree Model and Neural Network Model. The efficiency of these techniques is compared through sensitivity, specificity and accuracy. The existing datasets of Heart Disease patients from Cleveland Database of UCI repository is utilized to test and legitimate the execution of various calculations.

Key words: Heart disease Prediction, Logistic Regression, Neural Network, Random Forest

I. Introduction

Heart Disease Prediction, has always been the most fascinating and testing subject for medicinal analysis. According to the information from WHO, 33% populace overall kicked the bucket from coronary illness. Since, last decades, there is continuous evolution in Heart Disease Prevention and Treatment. Healthcare has a vast amount of data stored in it. The captured data have many important uses. Medical Professionals and researchers track the activity of hospitals and admittance of patients, facilities engaged on them etc. Mostly the health state and various factors, causing the disease are recorded. The data present is a collection of variables such as age, sex, obesity and various parameters adding to diseases in the simple numerical and categorical form. However, in this study we focus only on disease prediction, but there lie many facets of medical industry that can be improved and revolutionized through the intelligent use of machine learning techniques over data. The two term have been of concerned always, “prevention “and “treatment” that how to prevent and secure disease. But with the advent of new technology such as Machine Learning, there emerges a new term “Prediction”, which emphasis on prediction of future results from existing past encounters. Here, arises a new term” Predictive Analytics”, building predictive models using machine learning methods and analyzing the result with best of the available ability of Machine Learning tools, data Visualization. Prediction helps us to learn from past experiences and make a way to improve and enhance future results. Machine Learning methods have become the extreme fascinating and popular tools for medical researchers. Machine learning has enabled and enhanced the new ways of discovering and identifying patterns and finding complex relationships amongst them, from complex datasets. With the developing pattern of Machine Learning applications, for example, foreseeing re-permission consider, centrality of customized pharmaceutical framework and so forth. We present a Heart Disease Prediction model that predicts the probabilistic chances of having Heart Disease in future. To build a predictive model, Logistic Regression Model, Neural Network and Random Forest Model are used, analyzing their performances through comparative analysis of important evaluation parameters. The main concern in this study emphasizes on, which algorithm results chances with highest accuracy. R Language is being utilized for the current study. R is a powerful comprehensive statistical platform, which offers all manner of data analytics.

This paper proposes Heart Disease Prediction Model utilizing three of the best information mining strategies as Logistic Regression, Neural Network and Random Forest Model implemented in R Language.

II. Methodology

For building Predictive Model, R Studio is used which is a free and open source incorporated improvement condition (IDE) for R.

The basic steps used in R Scripts are:
1) Load the data.
2) Divide the data into train/test set.
3) Train the Model.
4) Test.
5) Evaluate on accuracy.
6) Save the results.

For comparing different techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/Heart+Disease. The dataset has 76 attributes and 303 records. From which, only 13 are selected.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0=female 1=male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang	Discrete	Exercise induced angina 1= Yes 0= No

Thalach	Continuous	Maximum Heart Rate Achieved
Old peak	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Heart	Discrete	0= No Heart Disease 1= Presence of Heart Disease

Table 1: Selected Attributes

A. *Logistic Regression:*

Logistic Regression is used when the response variable y is a binary variable, and autonomous factors or predictors are numeric. It models the connection amongst x and y by fitting a calculated logistic bend, which looks like S-Shaped or Sigmoid Curve. It fits a regression curve

$$y=f(x) \quad (1.1)$$

when y is in form of probabilities, or binary (“0” or “1”,” failure” or “success”). It finds application in such examples to predict, who will the next IPL match, next World Cup based on their quality and strength, or whether a person have disease or not (Yes or no).

B. *Neural Network:*

Artificial Neural Network are the networks of neurons, which is a model derived from neural structure of human brains. It is a monstrous parallel processor of interconnected neurons. A neuron consists of a set of information qualities, with associated weights and an input function that models the connection between input and output signal by processing through arithmetic or logical operations. The weights allow each of n inputs, (x), to contribute to a greater or lesser amount to the sum of input signals. The net total is used by the activation function f(x), and the resulting signal, y(x), is the output axon, forms an equation as follows:

$$(\varkappa)=f(\sum wi\varkappa i) \qquad (1.2)$$

C. *Random Forest:*

One of the best supervised learning approach is tree based modelling. Random Forest is one of the sort of tree based learning calculations that empowers predictive model with model stability and high accuracy. They are definitely not only adaptable to classification problems but they stand for regression also. Due its ability to ensemble, which involves grouping of predictive models, its able to achieve a better accuracy with stability. Random Forest are combination of decision tree predictors where each tree relies upon the independently sampled, random variable with same distributions for each tree in a forest. Random Forest overcome the problem of high variance and high bias, of a single decision tree. It also performs data exploration activities from missing worth treatment to feature engineering.

III. RESULT ANALYSIS

The execution of Algorithms is analyzed by assessing the sensitivity, specificity, accuracy and total time taken to execute. The Evaluation Parameters are estimated using “hmeasure” package in R. The hmeasure package is known for providing all measures of classification performance.

The sensitivity or affectability is a degree of positive instances that are correctly encountered as positive (i.e. the probability of patients known to have the ailment, who test positive for it). The specificity is the measure of negative instances that are correctly encountered as negative (i.e. the probability of patients known not to have the ailment, who test negative for it). The accuracy is the proportion of instances that are correctly classified. To quantify the dependability of the execution of proposed model, ROC curve is plotted. In statistics, a receiver operating characteristics curve, or ROC curve, is a primary tool for determining “cause and effect” relationship. It marks the tradeoff amongst sensitivity and specificity under different threshold values. It is used for evaluating performances of binary classifier. The bend is made by plotting the genuine positive rate (sensitivity) at y-hub against the false positive rate (1-specificity) at x-hub different limit settings.

ROC analysis helps to geek the optimal model from other suboptimal models for prediction, independent of other factors such as cost context and class distributions.

The closer the curve is to leftmost up corner, more the model is termed to be perfect.

The diagonal speaks to the limit settings. The dotted lines in below Figures represent the arched frame of the ROC curve (ROCH) for each classifier, i.e., the least convex curve that lies above its ROC curve.

Algorithm	Sensitivity	Specificity	Accuracy	Time(msec)
Logistic Regression	0.8	0.861	83.22%	10.42
Neural Network	0.821	0.902	86.58%	184.36
Random Forest	0.844	0.871	85.91%	152.75

Table 2: Comparison of Algorithms

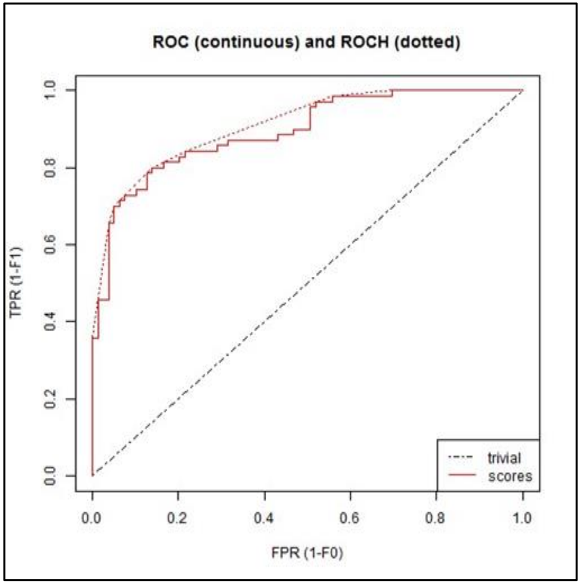

Fig. 1: ROC Curve for Logistic Regression

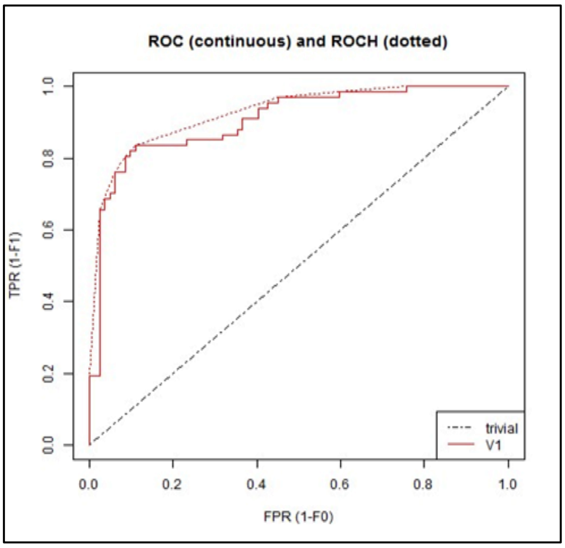

Fig. 2: ROC Curve for Neural Network

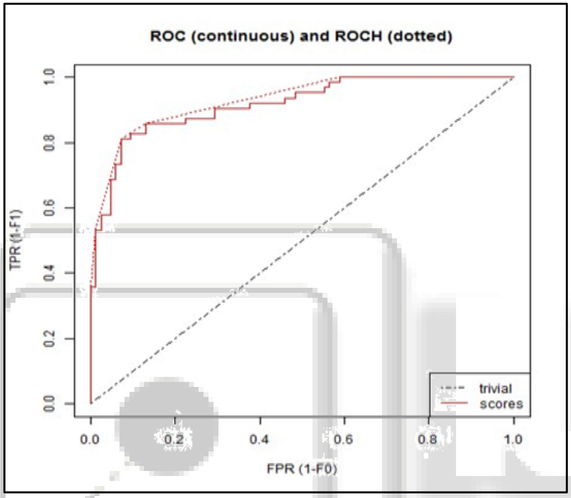

Fig. 3: ROC Curve for Random Forest

When comparing the result, Random forest has achieved highest True Positive Rate (Sensitivity) as 84% and Specificity, 87% less than Neural Network, 90% and minimum time from Neural Network. But Neural Network has achieved highest accuracy, 86.58% amongst the two with maximum time to execute. Logistic Regression is fastest model with 10.42 millisecond.

IV. Conclusion

By investigating the test comes about, it is inferred that Neural Network ended up being best classifier for Heart Disease Forecast. We can clearly see that highest accuracy belongs to Neural Network algorithm followed by Random Forest algorithm and Logistic Regression respectively. Also, observed that Neural Network has taken maximum time to build. Logistic Regression is fastest model with 10.42 millisecond.

In conclusion, we believe only a marginal success is accomplished in the production of prescient model for coronary heart disease patients and henceforth there is a requirement for effective and more perplexing models to increase the accuracy of predicting the early onset of heart disease.

References

[1] Wu R, Peters W, Morgan MW.The next generation clinical decision support: linking evidence to best practice. J Healthc Inf Manag, 2002; 16:50-5.

[2] Thuraisingham BM. A Primer for Understanding and applying data mining. IT Professional 2000; 1:28-31.

[3] Rajkumar A, Reena GS. Diagnosis of heart disease using datamining algorithm. Global Journal of Computer Science and Technology 2010; 10:38-43.

[4] Anbarasi M, Anupriya E, Iyengar NCHSN. Enhanced prediction of heart Disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology 2010; 2:5370-76.

[5] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. International Journal of Computer Science and Network Security 2008; 8:343-50.

[6] J. Han, M. Kamber, Data Mining: Concepts and Techniques, 2nd Edition, Morgan Kaufmann, 2006.