IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Comparative Study of Classification Methods using Algorithms of Data Mining for Possibilities of Heart Diseases

Bindu Trikha[1], Dhruv Dixit[2*]

[1,2]I.T. Department, IMS Ghaziabad (University Courses Campus), Ghaziabad, Uttar Pradesh

Corresponding Author: dhruv.dixit@imsuc.ac.in

DOI: https://doi.org/10.26438/ijcse/v7i6.330336 | Available online at: www.ijcseonline.org

Accepted: 09/Jun/2019, Published: 30/Jun/2019

Abstract: Rate of heart-related diseases are growing day by day on a greater pace from the last 15 years, which is a major concern. Through the classification, we can understand the possibilities of such worst-case scenarios at an earlier stage which can help us in being cautious and moreover being prepared for it in the immediate future.

***Keywords*:** Classification, possibilities of heart diseases, comparative analysis of algorithms using R.

I. INTRODUCTION

In recent years there is a rapid growth in the occurrence of heart diseases, and the same is witnessed at an early age than normal age of the presence of such diseases, which is a major problem. Now if we look upon the following researches by "World Health Organisation" we found that 17.9 million people die each year from CVD, an estimated 31% of all deaths worldwide.

So, these are the primary areas we need to focus to do the classification, so that we can ultimately act upon them by seeing categorized flaws observed during the process. Now, basically a classification can be done by various methods, but considering in mind the delicacy of the concern I have selected the top three best out of all Algorithms which provides the comparative analysis over the topic, which majorly shows the various aspects observed during the flow of algorithms and what are the consequences that we will get at the end by comparing all three of them. Basically, these algorithms are considered to make our research more precise and moreover in the amount of three, so it shows a clear comparison and most importantly the analysis and productive results for the greater good of the concerned people.

1. Algorithms selected for comparison

C5.0 Decision Tree, Random Forest, Naïve Bayes are the classification algorithms proposed in recent years. All of them have been adapted from current data mining and machine learning algorithms.

C5.0 Decision Tree

While there are numerous implementations of decision trees, one of the most well-known is the **C5.0 algorithm**. The C5.0 algorithm has become an industry standard for producing decision trees because it does well for most types of problems directly out of the box. Compared to more advanced and sophisticated machine learning models (e.g. Neural Networks and Support Vector Machines), the decision trees under the C5.0 algorithm generally perform nearly as well but are much easier to understand and deploy. One of the benefits of the C5.0 algorithm is that it is opinionated about pruning; it takes care of many of the decisions automatically using fairly reasonable defaults. Its overall strategy is to *post prune* the tree. It does this by first growing a large tree that overfits the training data. Afterward, nodes and branches that have little effect on the classification errors are removed.

II. RANDOM FOREST

It is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms because it's simplicity and the fact that it can be used for both classification and regression tasks. In this post, you are going to learn, how the random forest algorithm works and several other important things about it. Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The "forest" it builds, is an ensemble of Decision Trees, most of the time trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. One big advantage of random

forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Bayes' Theorem

This lets us examine the probability of an event based on the prior knowledge of any event related to the former event. So, for example, the probability that price of a house is high can be better assessed if we know the facilities around it, compared to the assessment made without the knowledge of the location of the house. Bayes' theorem does exactly that.

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)}$$

Above equation gives the basic representation of the Bayes' theorem. Here A and B are two events and,

P(A|B): the conditional probability that event A occurs, given that B has occurred. This is also known as the posterior probability.

P(A) and P(B): probability of A and B without regard to each other.

P(B|A): the conditional probability that event B occurs, given that A has occurred.

2. Software used for comparison:

We have done our experiments with C5.0 Decision Tree, Random Forest and Naïve Bayes Algorithm with R Language. Default settings are used for all compared ensemble methods. We were aware that the accuracy of some methods on some data sets can be improved when parameters were changed. However, it was difficult to find another uniform setting good for all data sets. Therefore, we did not change default settings since the default produced high accuracy on average.

3.Experimental Results & Discussion:

Implementation of Decision Tree on the Data Set:

> model_c50=C50::C5.0(heart_train[,-14],heart_train[,14])

> model_c50

Call:

C5.0.default(x = heart_train[, -14], y = heart_train[, 14])

Classification Tree

Number of samples: 212

Number of predictors: 13

Tree size: 14

Non-standard options: attempt to group attributes

> summary(model_c50)

Call:

C5.0.default (x = heart_train[, -14], y = heart_train[, 14])

C5.0 [Release 2.07 GPL Edition] Sat Jun 08 06:13:37 2019

III. DATA SET USED

Context: This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4.

Content

Attribute Information:

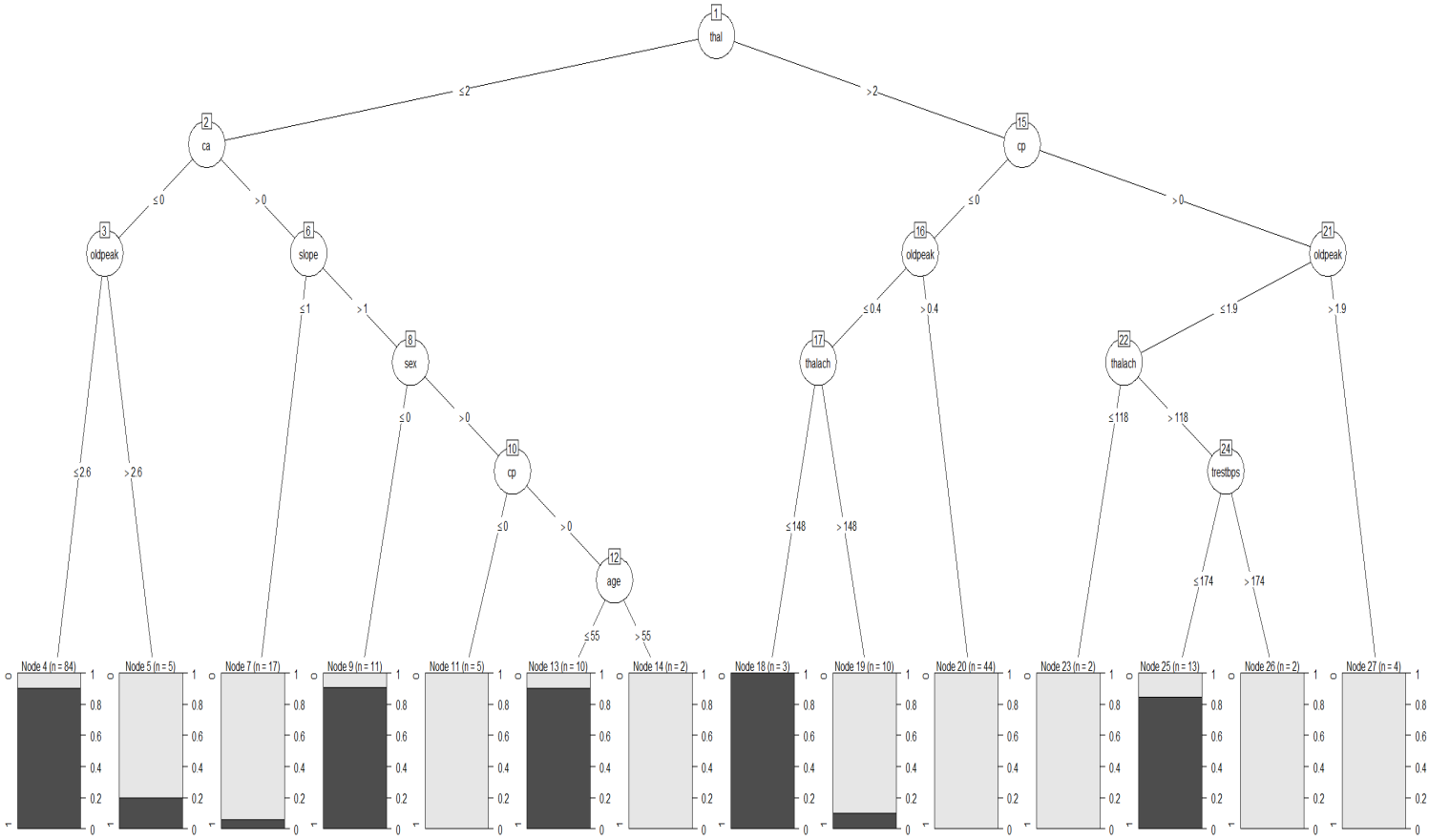
1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholesterol in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. Exercise-induced angina
10. old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) coloured by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect

Acknowledgments

Creators:
1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D


```
Class specified by attribute `outcome'

Read 212 cases (14 attributes) from undefined.data

Decision tree:

thal > 2:
:...cp <= 0:
:   :...oldpeak > 0.4: 0 (44)
:   :   oldpeak <= 0.4:
:   :   :...thalach <= 148: 1 (3)
:   :       thalach > 148: 0 (10/1)
:   cp > 0:
:   :...oldpeak > 1.9: 0 (4)
:       oldpeak <= 1.9:
:       :...thalach <= 118: 0 (2)
:           thalach > 118:
:           :...trestbps <= 174: 1 (13/2)
:               trestbps > 174: 0 (2)
thal <= 2:
:...ca <= 0:
    :...oldpeak <= 2.6: 1 (84/8)
    :   oldpeak > 2.6: 0 (5/1)
    ca > 0:
    :...slope <= 1: 0 (17/1)
        slope > 1:
        :...sex <= 0: 1 (11/1)
            sex > 0:
            :...cp <= 0: 0 (5)
                cp > 0:
                :...age <= 55: 1 (10/1)
                    age > 55: 0 (2)


Evaluation on training data (212 cases):

                Decision Tree
                ----------------
          Size      Errors

            14   15( 7.1%)   <<


           (a)   (b)    <-classified as
          ----  ----
            88    12    (a): class 0
             3   109    (b): class 1


Attribute usage:

100.00%thal
 78.77% oldpeak
 63.21% ca
 44.81% cp
 21.23% slope
 14.15% thalach
 13.21% sex
  7.08%  trestbps
  5.66%  age

Time: 0.0 secs

> predict_c50=predict(model_c50,heart_test)
> predict_c50
[1] 1 0 0 0 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 0 0 0 1 0 1 1 1 1 0
0 1 1 1 0 0 1 0 1 1 1 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0 0 1
[59] 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 0 1 0 0
1 0 0 1
```


Levels: 0 1



Implementation of Random Forest:

Random forest

>heart_forest=randomForest(target~.,data=heart_train)

> heart_forest

Call:

randomForest(formula = target ~ ., data = heart_train)

Type of random forest: classification

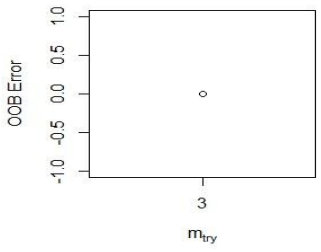
Number of trees: 500

No. of variables tried at each split: 3

OOB estimate of error rate: 17.92%

Confusion matrix:

	0	1	class.error
0	77	23	0.2300000
1	15	97	0.1339286



> heart_forest$importance

	MeanDecreaseGini
age	10.057967
sex	3.788342
cp	13.475079

trestbps 8.280866

chol 8.135939

fbs 1.133406

restecg 2.079026

thalach 11.933479

exang 4.254238

oldpeak 12.456453

slope 4.152575

ca 11.622739

thal 12.699571

> pred_heart=predict(heart_forest,newdata = heart_test,type="class")

> pred_heart

74 215 285 234 293 273 128 299 15 80 54 116 127 210 96 38 124 252 147 17 154 235 187 49 191 29 112 123 95 192

1 0 0 0 0 1 1 0 1 1 1 1 1 0 0 1 1 0 1 1 1 1 0 0 1 0 1 1 1 1 0

102 79 44 77 247 218 3 169 159 12 10 48 211 105 103 270 233 241 9 200 180 28 78 176 194 171 280 82 263 13

0 1 1 1 0 0 1 0 0 1 1 1 0 1 1 0 0 0 1 0 0 1 1 0 0 1 0 1 0 1

33 229 250 150 134 58 71 281 97 256 84 2 34 300 65 50 173 294 4 21 8 88 189 107 230 261 114 20 202 248

1 1 0 1 1 1 1 0 1 0 1 0 1 0 1 1 1 1 1 0 1 1 1 0 0 0 0 1 0 0

24

1 Levels: 0

Implementation of Naïve Bayes :

Naïve Bayes Classification

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

>model_nav=naiveBayes(target~.,data=heart_train)

> model_nav

Naive Bayes Classifier for Discrete Predictors

A-priori probabilities:

Y

0 1

0.4716981 0.5283019

Conditional probabilities:

age

Y [,1] [,2]

0 56.04000 8.015036

1 52.55357 10.208493

sex

Y [,1] [,2]

0 0.8100000 0.3942772

1 0.5267857 0.5015260

cp

Y [,1] [,2]

0 0.4400 0.8798531

1 1.3125 0.9775060

trestbps


```
Y       [,1]     [,2]
 0 133.6600 19.50976
 1 128.3304 15.85258


  chol
Y       [,1]     [,2]
 0 247.8500 50.17170
 1 243.1964 54.21094


  fbs
Y       [,1]      [,2]
 0 0.1500000 0.3588703
 1 0.1071429 0.3106849


  restecg
Y       [,1]      [,2]
 0 0.4900000 0.5594911
 1 0.6160714 0.5066323


  thalach
Y       [,1]     [,2]
 0 140.8100 22.92090
 1 157.9375 19.93199



  exang
Y       [,1]      [,2]
 0 0.4900000 0.5024184
 1 0.1517857 0.3604257


  oldpeak
Y       [,1]     [,2]
 0 1.6520000 1.411960
 1 0.6044643 0.735997


  slope
Y       [,1]      [,2]
 0 1.200000 0.6030227
 1 1.580357 0.5948486


  ca
Y       [,1]      [,2]
 0 1.1300000 1.0115994
 1 0.4017857 0.9050392


  thal
Y       [,1]      [,2]
 0 2.550000 0.6571287
 1 2.080357 0.4274035


> predict_nav=predict(model_nav,heart_test)
> predict_nav
 [1] 1 0 0 0 0 0 1 0 1 0 1 1 1 0 0 1 1 0 1 1 1 0 0 1 0 1 1 1 1
0 0 1 1 0 0 0 1 0 0 1 1 1 1 1 1 0 0 0 0 1 0 1 1 0 0 0 0 1 0 1 1
1 0 1 1 1 1
[68] 0 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 1 0 0 0
Levels: 0 1
```

IV. COMPARISON

	C5.0 Decision Tree	Random Forest	Naïve Bayes
Accuracy	0.7802	0.8571	0.8242
95% CI	(0.6812, 0.8603)	(0.7681, 0.9217)	(0.7302, 0.896)
No. of Information Rate	0.5824	0.5824	0.5824
P-Value [Acc > NIR]	5.82E-05	1.42E-08	7.69E-07
Kappa	0.5413	0.7074	0.6412
Mcnemar's Test P-Value	0.5023	1	0.8026
Sensitivity	0.6842	0.8421	0.8158
Specificity	0.8491	0.8679	0.8302
Pos Pred Value	0.7647	0.8205	0.775
Neg Pred Value	0.7895	0.8846	0.8627
Prevalence	0.4176	0.4176	0.4176
Detection Rate	0.2857	0.3516	0.3407
Detection Prevalence	0.3736	0.4286	0.4396
Balanced Accuracy	0.7666	0.855	0.823

V. CONCLUSION

In this paper, we have presented an intelligent and effective heart disease prediction methods using data mining. We studied an efficient approach for the extraction of significant patterns from the heart disease data warehouses for the efficient prediction of heart disease. Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. Real data from Health care organizations and agencies needs to be collected and all the available techniques will be compared for the optimum accuracy.

In this study, C5.0 Decision Tree, Random Forest, and Naïve Bayes were implemented on a Heart Diseases Dataset to predict the potential risk in the future. Based on the three types of scenario results, Random Forest achieves better performance. It clearly states that the highest Balanced Accuracy is of Random Forest for the following Data Set, so it is preferable to use Random Forest for this type of Data set. Whereas the Naïve Bayes shows slightly better in the corresponding terms to C5.0 Decision Tree.

Moreover, can conclude that the presence of heart diseases can be predicted through such methods which help us in being aware and analyzing the stuff in a more efficient manner.

REFERENCES

[1] C. S. Dangare, S S. Apte Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques , International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[2] J . Soni, U. Ansari, D. Sharma, S. Soni. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications (0975 – 8887)Volume 17– No.8, March 2011

[3] K. Srinivas, B. Kavihta Rani, Dr A. Govrdhan. Application of data mining techniques in healthcare and prediction of heart attacks .

Authors Profile

Mr. Dhruv Dixit is pursuing Bachelor of Computer Application from **IMS Ghaziabad (University Courses Campus)** since 2017.

Ms. Bindu Trikha is working as an Assistant professor, **IMS Ghaziabad University courses Campus**. She has an approximate 15 years of academic experience. Her areas of interest are Data analysis, Data mining & Algorithm designing.